INTERACTIVE CONTROL ENGINEERING

Control in Semiconductor Manufacturing

Interactive Control Engineering: Principles and Applications

Edited by John D. Hedengren

Control technology

1 Process Dynamics, Modeling, and Identification
2 Practical PID Control
3 Feedforward, Cascade, and Ratio Control
4 Override, Split-Range, and Constraint Control
5 Delays, Slow Measurements, and State Estimation
6 Multivariable and Plantwide Control
7 Nonlinear Dynamics, Oscillations, and Stabilization
8 Batch, Run-to-Run, and Hybrid Control
9 Model Predictive Control and Economic Optimization

Industry applications

10 Control in Upstream and Midstream Oil and Gas
11 Control in Petroleum Refining
12 Control in Petrochemicals and Polymers
13 Control in Chemical Plants
14 Control in Semiconductor Manufacturing
15 Control in Mining and Mineral Processing
16 Control in Iron, Steel, and Metals Production
17 Control of Energy Conversion and Storage
18 Control of Nuclear and Fusion Systems
19 Control of Biological Systems and Human Performance
20 Motion Control, Robotics, and Autonomous Systems

INTERACTIVE CONTROL ENGINEERING
Principles and Applications

Volume 14

Control in Semiconductor Manufacturing

Run-to-run control where little can be measured in the chamber

John D. Hedengren
Editor

Advanced Process Solutions, LLC

Interactive Control Engineering: Principles and Applications
Volume 14: Control in Semiconductor Manufacturing
Edited by John D. Hedengren

First edition, 2026

Copyright © 2026 Advanced Process Solutions, LLC. All rights reserved.

Produced by Advanced Process Solutions, LLC.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means without the prior written permission of the publisher, except for brief quotations in a review or scholarly work.

The simulators described in this book are teaching models. Their equations, parameters and results are simplified for instruction and are not validated for the design, operation or safety assessment of any real plant, vehicle or medical device. The editor and the publisher accept no liability for use of this material outside that purpose.

Names of products and companies are used for identification only and may be trademarks of their owners.

Every chapter has a companion simulator that runs in a web browser at https://apmonitor.com/apps/. This volume online, with chapter files and citation records: https://apmonitor.com/apps/books/v14-semiconductor/

How to cite this volume. John D. Hedengren, Ed., *Control in Semiconductor Manufacturing*, vol. 14 of *Interactive Control Engineering: Principles and Applications*. Advanced Process Solutions, LLC, 2026.

Typeset in Times and Helvetica-family type with LaTeX and the memoir class.

Preface

This book is one of 20 volumes of *Interactive Control Engineering: Principles and Applications*. Each chapter is about one plant, machine or system and one control lesson, and each has a simulator that runs in a web browser with nothing to install. The simulators are collected at https://apmonitor.com/apps/. Nine volumes are organized by control technology and eleven by industry or application area; the last pages of this volume list them all.

How to read this book with the apps

Chapter 1 teaches the theory this volume relies on and explains, once, the simulator framework that the apps share. Each chapter after it opens with the control question it answers, its learning objectives, what to read first, and the address of its app, also given as a code to scan. The chapter then develops the model, the controllers and the measured results. Every number reported was measured on the app, so a reader can open the app, load the named scenario and reproduce it.

Every chapter ends with exercises. Each is labeled with its kind and the time it should take, and none takes more than ten minutes. Concept checks and calculations can be done on paper. App challenges name a scenario, the settings and the run length. One exercise in each chapter, labeled *Review the agent*, presents a tuning, a model or a claim proposed by an AI assistant and asks what is wrong with it. Short answers to the odd-numbered exercises are at the back of the book, and worked solutions are available to instructors.

How AI agents were used

AI agents did much of the development work behind this series. Under the editor's direction they wrote and tested the simulators, drafted the papers from which the chapters are drawn, measured the reported results on the running apps, converted the papers to book form, and drafted the introductory and closing chapters and the exercises. Other agents then checked that work: they re-measured reported numbers, audited what each app displays against the state of its model, reviewed the papers as referees, and solved the exercises without seeing the answers. The editor set the scope, chose the plants and the lessons, reviewed the results and is responsible for the content.

That way of working is also a subject of the book. The last chapter of each volume describes what the agents got wrong while this volume was made and how the errors were caught, and it sets out what the editor believes this means for engineers: more of the routine work can be delegated, and more of the engineer's value lies in checking, supervising, choosing the problem and knowing the process.

About this volume

This volume follows a wafer through semiconductor manufacturing, from polysilicon and crystal growth through lithography, bake, etch, implant, anneal, plating and polish. Little can be measured inside a process chamber, and the wafer is measured after it has left. The chapters show how each tool is held on target by feedforward from the previous step, by fast loops on what can be measured, and by correction from one run to the next through a metrology queue.

Each chapter names the technology volume that teaches its main technique. Volume 8 develops run-to-run control on other batch processes.

John D. Hedengren
2026

Contents

Preface v

Notation x

1 Introduction: Late Measurement and Run-to-Run Correction 1
- 1.1 The Question This Volume Answers . . . 1
- 1.2 A Short History . . . 2
- 1.3 Notation and Conventions . . . 4
- 1.4 Measuring Late and Correcting the Next Wafer . . . 4
- 1.5 Instruments and Final Elements in This Volume . . . 12
- 1.6 The Simulator Framework . . . 14
- 1.7 Working With AI Agents While Studying . . . 17
- 1.8 Chapter Roadmap . . . 18
- 1.9 Exercises . . . 21

2 Siemens Polysilicon Reactor Simulator for Gain Scheduling and an Unmeasured Constraint 23
- 2.1 Introduction . . . 24
- 2.2 Polysilicon Production and Its Instrumentation . . . 25
- 2.3 Reactor Model . . . 26
- 2.4 Three Properties of the Plant . . . 32
- 2.5 Controller Design . . . 34
- 2.6 Results . . . 37
- 2.7 Technology in Practice . . . 44
- 2.8 Limitations . . . 46
- 2.9 Discussion . . . 46
- 2.10 Conclusion . . . 47
- 2.11 Exercises . . . 49

3 Czochralski Crystal Puller Simulator for Mid-Ranging a Fast Limited Actuator 52
- 3.1 Introduction . . . 53
- 3.2 Automatic Diameter Control . . . 54
- 3.3 Puller Model . . . 55
- 3.4 Dynamic Analysis . . . 60
- 3.5 Control Structure . . . 64
- 3.6 Results . . . 66
- 3.7 Technology in Practice . . . 73
- 3.8 Limitations . . . 75
- 3.9 Discussion . . . 76
- 3.10 Conclusion . . . 77
- 3.11 Exercises . . . 78

4 Wafer Scanner Simulator for Preview Feedforward Where No Measurement Exists 82
- 4.1 Introduction . . . 83
- 4.2 From Contact Printing to the Scanning Slit . . . 84
- 4.3 Process Model . . . 85
- 4.4 Why the Preview Distance Belongs to the Scan Speed . . . 91
- 4.5 Implemented Controllers . . . 93

4.6 Results . . . 94
4.7 Technology in Practice . . . 100
4.8 Limitations . . . 103
4.9 Discussion . . . 103
4.10 Conclusion . . . 104
4.11 Exercises . . . 105

5 A Hundred-Zone Bake Plate Simulator for Decentralized Control and Spatial Bandwidth 109
5.1 Introduction . . . 110
5.2 Background . . . 111
5.3 Process Model . . . 112
5.4 The Spectrum of the Plate . . . 116
5.5 Controller Design . . . 118
5.6 Results . . . 121
5.7 Technology in Practice . . . 129
5.8 Limitations . . . 130
5.9 Discussion . . . 131
5.10 Conclusion . . . 132
5.11 Exercises . . . 133

6 Lithography and Etch Virtual Laboratory for Run-to-Run Control Through a Metrology Queue 137
6.1 Introduction . . . 138
6.2 From Control Charts to Recipe Feedback . . . 139
6.3 Process Model . . . 140
6.4 Controllers as Implemented . . . 144
6.5 Stability Through the Queue . . . 147
6.6 Results . . . 150
6.7 Technology in Practice . . . 155
6.8 Limitations . . . 157
6.9 Conclusion . . . 158
6.10 Exercises . . . 160

7 A Plasma Etch Simulator for Recipe Trades and Endpoint Detection 163
7.1 Introduction . . . 164
7.2 Background . . . 165
7.3 Interaction Analysis . . . 166
7.4 Chamber Model . . . 167
7.5 Controller Design . . . 169
7.6 Results . . . 172
7.7 Technology in Practice . . . 179
7.8 Limitations . . . 181
7.9 Discussion . . . 182
7.10 Conclusion . . . 183
7.11 Exercises . . . 184

8 An Ion Source Simulator for Cascade Control and Recovery From a Measurement That Stops Meaning Anything 188
8.1 Introduction . . . 189
8.2 Sources, Sparks and Held Outputs . . . 190
8.3 Process Model . . . 191
8.4 Controllers as Implemented . . . 197

8.5 Two Judgments That Are Usually One . . . 198
8.6 Results . . . 199
8.7 Technology in Practice . . . 204
8.8 Limitations . . . 206
8.9 Conclusion . . . 206
8.10 Exercises . . . 208

9 Rapid Thermal Processing Simulator for Ramp Feedforward and Pyrometer Emissivity Error 212
9.1 Introduction . . . 213
9.2 Lamps, Pyrometers and the Spike Anneal . . . 214
9.3 The Chamber as Modeled . . . 215
9.4 What Radiation Does to the Loop . . . 220
9.5 Implemented Controller . . . 221
9.6 Measured Behavior of Each Case . . . 225
9.7 Technology in Practice . . . 233
9.8 Limitations . . . 235
9.9 Conclusion . . . 235
9.10 Exercises . . . 237

10 Copper Plating Bath Simulator Dosed by Amp-Hours and Analyzed Every Two Hours 240
10.1 Introduction . . . 241
10.2 Brighteners, Stripping Analysis and Damascene Fill . . . 242
10.3 The Bath as Modeled . . . 243
10.4 A Window Narrowed by an Inventory . . . 247
10.5 Implemented Controller . . . 249
10.6 The Gain a Two-Hour Analysis Can Carry . . . 251
10.7 The Cases as Measured . . . 253
10.8 Technology in Practice . . . 258
10.9 Limitations . . . 259
10.10 Conclusion . . . 260
10.11 Exercises . . . 262

11 A Chemical Mechanical Planarization Simulator for Profile Control Across a Consumable Lifecycle 265
11.1 Introduction . . . 266
11.2 From Pad Wear to Consumable-Aware Recipe Control . . . 267
11.3 Process Model . . . 268
11.4 Controllers as Implemented . . . 273
11.5 What a Maintenance Event May Delete . . . 276
11.6 Results . . . 277
11.7 Technology in Practice . . . 281
11.8 Limitations . . . 284
11.9 Conclusion . . . 285
11.10 Exercises . . . 286

12 Summary and Outlook 290
12.1 What the Chapters Showed . . . 290
12.2 Choosing a Method . . . 293
12.3 Where Practice Is Heading . . . 294
12.4 Engineering With AI Agents . . . 296
12.5 Gaps in Semiconductor Process Control . . . 298

12.6 Exercises 303

Answers to Selected Exercises **304**

Glossary **308**

Index of Control Topics **311**

Index of Applications **314**

About the Editor **316**

Volumes in This Series **317**

Notation

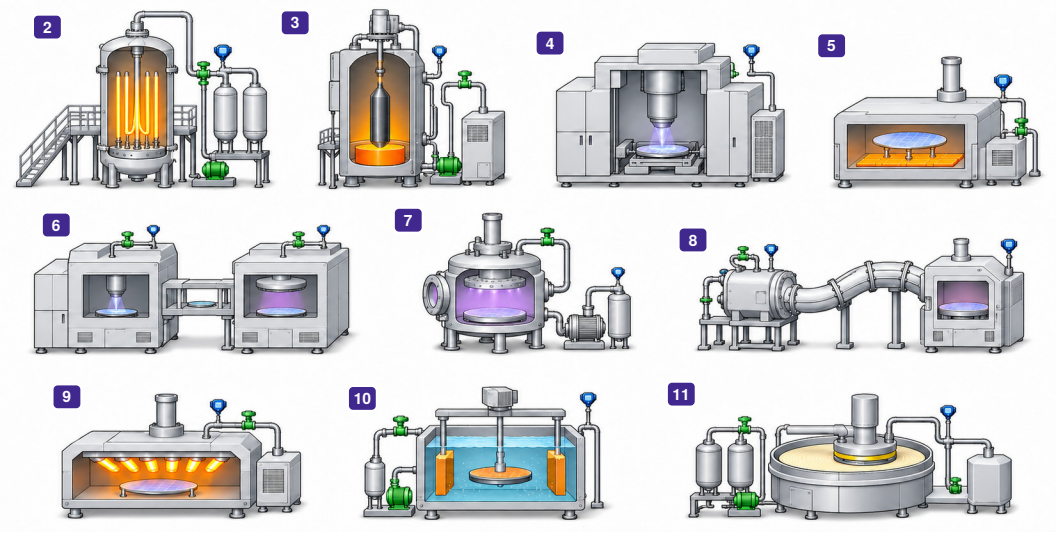
The chapters share the symbols below. A symbol used by one chapter only is defined where it first appears, and each chapter states its own units.

Symbol	Meaning
SP	setpoint, the value a controller is asked to hold
PV	process variable, the measurement a controller acts on
CV	controlled variable
MV	manipulated variable; also OP, the controller output
DV	disturbance variable
e	control error, SP $-$ PV
K_c	controller gain
τ_I, τ_D	integral (reset) time and derivative time
K_p	process gain, the steady-state change in PV per unit change in MV
τ_p	process time constant
θ_p	process dead time
τ_c	closed-loop time constant chosen in a tuning rule
Δt	controller sample time
s	Laplace variable
x, u, y, d	state, input, output and disturbance of a state-space model
A, B, C	matrices of the linear model $\dot{x} = Ax + Bu$, $y = Cx$
Q, R	weights on outputs (or states) and on input moves in an optimal controller
$\hat{x}$	an estimate of x

Abbreviations used throughout: PID, proportional-integral-derivative; PI, proportional-integral; FOPDT, first-order plus dead time; MPC, model predictive control; MHE, moving horizon estimation; IAE, integral of the absolute error; RGA, relative gain array. Times are in seconds unless a chapter says otherwise. Controller output and valve position are in percent of span. A scenario or case number refers to the numbered list in the chapter's app.

CHAPTER 1

Introduction: Late Measurement and Run-to-Run Correction



The chapters 2 polysilicon reactor 3 crystal puller 4 wafer scanner 5 bake plate 6 lithography and etch 7 plasma etcher 8 ion implanter 9 spike anneal 10 copper plating bath 11 polisher

Notice that the top row makes the silicon and prints the pattern, the middle row transfers it and dopes the wafer, and the bottom row anneals, wires and flattens it. Illustration, not to scale.

IN PLAIN TERMS **Correcting the Next Wafer From the Last One**

The main idea. A chip is built on a thin disk of silicon, called a wafer, in several hundred steps. Most steps leave nothing that can be measured while they run, so the wafer is checked afterward on another machine.

The challenge. That check often comes back after more wafers have been made, and only some wafers are checked at all. Meanwhile the machines change, as their walls get coated, their pads wear and their baths age.

Why it matters. Good control uses each late check to correct the wafers that follow, so that features far smaller than a hair come out the right size.

THREE WAYS TO READ THIS CHAPTER

Explore	8 min	This page, Section 1.2 with its timeline and two History boxes, Figure 1.2, the Did you know? boxes in Sections 1.4.1, 1.4.3 and 1.4.4, and Key points in Section 1.8. No equations.
Study	45 min	The whole chapter in order, working each number of the example M-1, then Exercises 1 to 10.
Practice	5 min	This page, Section 1.4.8 with Table 1.1, Section 1.5 with Figures 1.7 and 1.10, and Section 1.8.

1.1 The Question This Volume Answers

A chip is built on a silicon wafer in several hundred steps, and most of the steps leave nothing that can be measured while they run. The linewidth that an etcher cuts, the dose an implanter delivers and the copper a plating cell lays down are known only after the wafer has left the tool, on a sample of the wafers, and often after other lots have already been processed. The tools themselves drift between those measurements. Chamber walls season, pads wear, baths age and windows warm, and every few days a maintenance event replaces part of the plant. This volume answers one question: how is a step

controlled when the property that matters is measured afterward, late and on a sample, and which techniques recur from one tool to the next?

Seven control problems define the industry, and most chapters meet more than one of them.

- *The product is not measured where it is made.* Chamber sensors read pressure, power and temperature. The wafer's linewidth, thickness or sheet resistance is read on another tool.
- *Measurements arrive late and on a sample.* A reading of one lot returns after later lots have run, and many lots are not measured at all.
- *The tool changes between runs, or while it runs.* Seasoning, wear and aging move the plant from wafer to wafer, and a rod or a crystal that grows during the batch moves it within one run.
- *One wafer, many zones.* A plate, a lamp bank or a carrier head has many actuators that act on one continuous wafer and on each other's zones.
- *A step ends on an event.* An etch is stopped when the film clears, not when a clock runs out.
- *Errors pass from step to step.* A wide line after lithography stays wide after etch unless the etcher is told, so one step can correct the error of the step before.
- *A measurement can be wrong while it looks right.* A pyrometer that assumes the wrong emissivity, a gap voltage that has sparked or a metrology tool with an offset all give readings a loop will act on.

Ten chapters answer the question on ten tools, each with an app that the reader runs in a browser. Chapters 2 and 3 make the material: polysilicon rods that are their own heaters, and a single crystal pulled from the melt. Chapters 4 to 6 pattern it, with a scanner that focuses on a surface it cannot see under the slit, a hundred-zone bake plate and a lithography and etch line run lot by lot. Chapters 7 to 9 etch, implant and anneal, and Chapters 10 and 11 plate the copper wiring and polish it flat.

A reader who has worked through the volume can do the following on a tool that is new to the reader.

- Say what a tool measures, where and when, and which property of the wafer no instrument on it reports.
- Write a run-to-run controller, choose its weight against noise and drift, and find the weight at which a metrology queue makes it unstable.
- Explain why many well-tuned zone loops can overshoot together, and what a model of the zone coupling adds.
- Decide when a step should end on an endpoint signal instead of a timer, and how much over-etch the signal saves.
- Correct one step from a measurement of the step before, and say when the correction lowers the spread.

This chapter develops each of these on one small example and names the chapter that uses it. PID forms and tuning are taken as known from Volume 2, feedforward and cascade from Volume 3, mid-ranging and override from Volume 4, dead time and estimation from Volume 5, interaction and the relative gain array from Volume 6, gain scheduling from Volume 7, run-to-run and batch control in general from Volume 8 and predictive control from Volume 9.

1.2 A Short History

The milestones of Figure 1.1 follow the tools of this volume, from the crystal and the rods that make the silicon to the steps that pattern, etch, implant, anneal, plate and polish it, and to the controllers that correct each step from the one before.

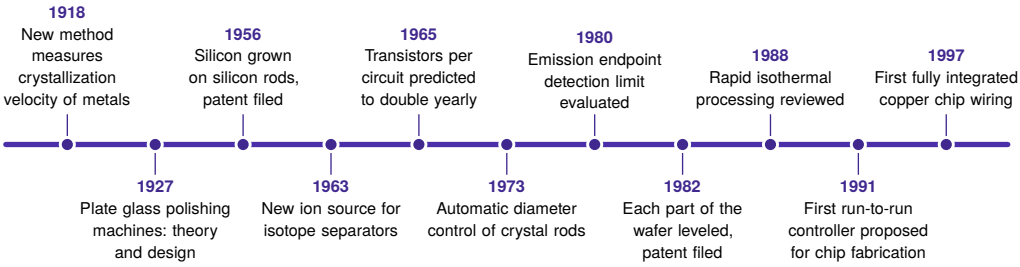


Figure 1.1 Milestones of semiconductor manufacturing and its control. The dates are from [1] (1918), [2] (1927), [3] (1956), [4] (1963), [5] (1965), [6] (1973), [7] (1980), [8] (1982), [9] (1988), [10] (1991) and [11] (1997). A patent is dated by its first filing. Notice that run-to-run control came after most of the tools it now corrects. Events are evenly spaced, not to a time scale.

> ⧖ HISTORY **1918 to 1973: Growing the Silicon**
>
> In 1918 Czochralski published a new method for measuring how fast metals crystallize [1], and the puller of Chapter 3 carries his name. A patent first filed in Germany in June 1956 grew pure silicon on self-supporting rods of silicon, joined at their upper ends so that the current passed up one rod and down the other. The rods' temperature was judged by their glowing color, and the current was raised by hand with a rheostat, because the resistance fell as the rods thickened [3]. A paper published in 1973 held the diameter of a pulled silicon crystal automatically. It weighed the crystal and moved the pull rate, and a second controller trimmed the heater power slowly, waiting and seeing, whenever the pull rate left a set band [6]. Chapter 2 and Chapter 3 control both.

The pace of everything after was set by a prediction. In 1965 Gordon Moore predicted, from a few data points, that the number of transistors on an integrated circuit would double about every year. The trend was later adjusted to a doubling every 18 to 24 months and has continued for more than half a century [5]. Polishing came from an older trade. A study of 1927 on plate glass polishing machines related the rate of removal to the pressure and to the relative velocity of pad and glass, and most models of chemical mechanical planarization are still inspired by it [2]. In 1963 a new ion source was described for electromagnetic isotope separators [4]. Some thirty years later it was still the most widely used ion source, and it was used extensively for the production doping of silicon devices [12].

> ⧖ HISTORY **1980 to 1995: Correcting From the Last Run**
>
> In 1980 a study proposed a measure for detecting the end of a plasma etch from the light of the discharge and evaluated its lower limit of detection for films of polycrystalline silicon and silicon nitride [7]. The step could now end on an event, but its result was still measured afterward. A run-to-run controller, which adjusts the recipe lot by lot or wafer by wafer, was first proposed for chip fabrication in 1991 [10]. A 1993 analysis gave a simple condition on the weight and the estimated process gain that keeps such a controller stable [13]. In 1995 a framework combined statistical process control with feedback. Its rapid mode answered sudden shifts such as those caused by maintenance, its gradual mode the slow build-up of deposits in a reactor, and on an epitaxy reactor the rapid mode recovered within three runs [14].

Patterning changed in the 1980s. Chemically amplified resists were proposed in the early 1980s, when exposure tools moved into the deep ultraviolet and their lamps gave too little light there. Rated at first a laboratory curiosity, the resist went quickly into the production of one-megabit memories [15]. A patent application filed in 1982 explained why a wafer could no longer be set square to the lens from three autofocus points. Larger wafers, new materials and each round of exposure and treatment made its surface uneven, so the height of each part of the wafer had to be detected [8]. In 1988 a

review called rapid isothermal processing, which heats one wafer with incoherent light, a superior alternative to the furnace, and it named more accurate temperature measurement as a need [9]. The wiring changed last. In 1997 the first fully integrated copper wiring of a chip technology stacked up to six levels of copper lines [11], the metal that Chapter 10 plates and Chapter 11 polishes flat.

1.3 Notation and Conventions

A *run* is one wafer or one lot, whichever the controller updates on, and k counts runs. A run-to-run controller applies the recipe u_k and later reads the result y_k against a target y^*. The process gain is b, the gain the controller believes is $\hat{b}$, and their ratio is $\xi = b/\hat{b}$. The weight of an exponentially weighted moving average (EWMA) is λ, and m is the number of runs that pass between a run and the arrival of its reading. Linewidths are in nanometers, temperatures in °C with differences in kelvin, and times in seconds or minutes as the tool runs. The framework's h is its integration step and its a the controller action, so neither is used for anything else here. Loop action follows Section 1.6.2: $a = +1$ is reverse acting and $a = -1$ direct acting. A recipe loop whose output rises when the result is above target, such as more trim for a wide line, is direct acting.

The chapters keep their own symbols, and several collide. λ is an EWMA weight in Chapter 6 (λ_E, λ_L), Chapter 5 (λ_z), Chapter 9 and Chapter 11, an analysis average in Chapter 10, and a set of regularization weights in Chapter 11. m is the metrology delay in Chapter 6 but an exponent in Chapter 9. θ is the chamber wall condition in Chapter 6, the analysis age in Chapter 10 and a vector of pad parameters in Chapter 11, while Chapter 7 writes its wall coverage as θ_w. ρ is a drift rate, a spectral radius, a ramp factor or a normalized radius, depending on the chapter. Chapter 3 and Chapter 10 write the action as γ, and Chapter 10 uses a for a decay factor $e^{-T/\tau}$, not for the action.

1.4 Measuring Late and Correcting the Next Wafer

1.4.1 The Wafer Flow

Figure 1.2 places each chapter's tool in the flow of a wafer [16]. Trichlorosilane deposits electronic-grade silicon on heated rods, and the rods are melted and pulled into a single crystal that is sliced and polished into wafers. Each layer of a chip is then patterned. A scanner exposes a resist, a bake plate finishes its chemistry and a developer opens the pattern, which an etcher transfers into the film beneath. In the front end an implanter places dopants and a lamp-heated chamber anneals them. In the back end a plating bath fills the etched trenches with copper and a polisher removes the excess. A wafer passes through the patterning steps once for every layer, and between the steps metrology tools measure a linewidth, a thickness or a sheet resistance on a sample of the wafers.

> **+ DID YOU KNOW? From Blue Light to Extreme Ultraviolet**
>
> To shrink the features, the light of lithography moved from blue at 436 nm to ultraviolet at 365, deep ultraviolet at 248 and 193, and extreme ultraviolet at 13.5 nm. Light that short is absorbed by every material, air included, so it is shaped only by mirrors coated with up to 100 layers of molybdenum and silicon [5].

The techniques the chapters share are developed below on one running example, a patterning module called M-1, drawn in Figure 1.3. A lot of 25 wafers enters every 30 min. A bake plate of five zones in a row, each with its own heater and thermocouple, bakes the exposed resist, and the developed linewidth is measured on every lot at once. An etcher trims the line for u seconds at 1.0 nm/s and then etches the film beneath, ending the main etch on a timer or on an optical endpoint. The finished linewidth, 40 nm on target, is measured two lots later. Every number of M-1 is declared here or computed from these declarations by the script kept with this chapter's record. The numbers are chosen for clear arithmetic and are not taken from any one chapter.

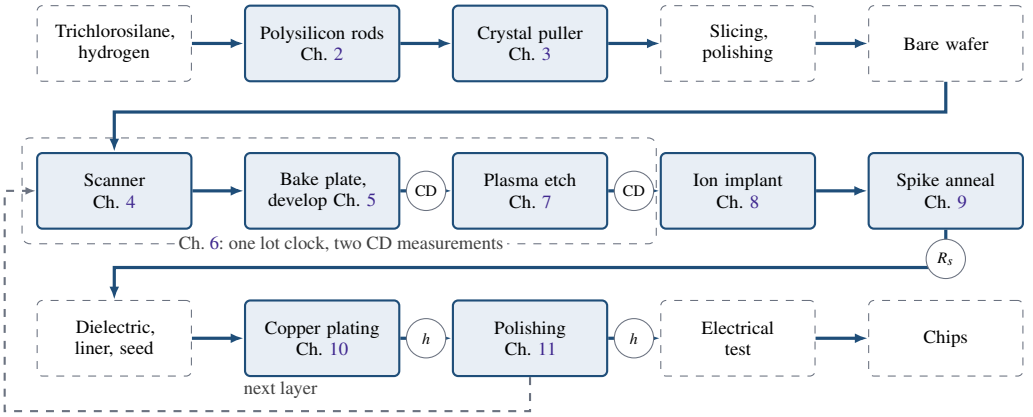


Figure 1.2 The Tools of the Volume in the Flow of a Wafer. Filled Blocks Are the Tools of Chapters 2 to 11, Dashed Blocks Lie Outside the Volume. Circles Mark Measurements on the Wafer After a Step: Linewidth (CD), Sheet Resistance (R_s) and Film Thickness (h). The Dashed Frame Is the Line of Chapter 6, and the Dashed Arrow Repeats the Patterning for the Next Layer.

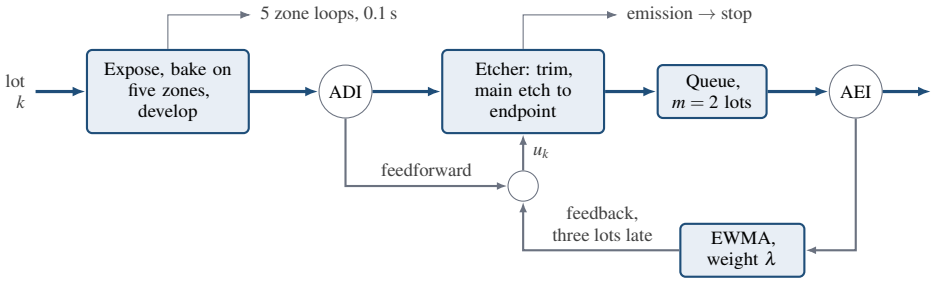


Figure 1.3 The Running Example M-1. The Developed Linewidth (ADI) Is Measured on Every Lot at Once and Can Correct the Etcher's Trim on the Same Lot. The Finished Linewidth (AEI) Is Measured Two Lots Later, so the Feedback Reaches the Trim on the Third Lot After the One It Describes. The Bake Plate Runs Five Zone Loops on a 0.1 s Scan, and the Etcher Ends Its Main Etch on an Emission Signal or a Timer.

1.4.2 Why Little Is Measured in the Chamber

The instruments of a process tool measure the tool. M-1's bake plate reads five thermocouples every 0.1 s and its etcher reads pressure, power and the light of the plasma. None of them reads the linewidth, because the line is tens of nanometers wide and forms under the resist, inside a plasma or under a slit of light that an instrument would block. The linewidth is read on a scanning electron microscope that serves many tools, on a sample of the wafers, after the lot has waited its turn [17]. Measurements therefore come in three kinds: tool sensors in the chamber that are fast and say little about the product, metrology after the step that is slow and says what the product is, and electrical test at the end of the line, weeks later, which no chapter uses.

The clocks of M-1 differ by large factors. One lot of 30 min is 18000 scans of a zone loop, and the after-etch reading of a lot changes the recipe of the third lot after it, 90 min after that lot was etched. A loop on a zone thermocouple corrects within seconds, and a loop on the finished linewidth corrects once a lot on old news. The chapters keep the same split. Chapter 7 closes four chamber loops that settle within 45 s, yet scores each wafer on seven criteria that only the finished wafer shows. Chapter 10 analyzes its bath every 2 h, and each result describes the bath of at least 42 min before. Chapter 11 receives each removal map one or two lots after the wafer was polished, and Chapter 9

measures sheet resistance only after the wafer has left the chamber.

Between metrology results, a model can estimate the product from the tool's own sensors, which the industry calls virtual metrology [18]. Chapter 2 infers the unmeasured center temperature of a growing rod from the supply's volts and amperes, Chapter 10 carries its breakdown products on a model corrected every 8 h by a laboratory, and Chapter 11 keeps a Bayesian estimate of a fresh pad. Estimation in general is the subject of Volume 5.

A fast measurement can also be wrong while it looks right. A pyrometer converts radiance to temperature with an emissivity it is given. Radiance from a wafer at absolute temperature T grows as εT^4, so a total-radiation pyrometer set to $\hat{\varepsilon}$ reads the temperature of Equation (1.1), and a loop that holds the reading on T_{SP} holds the wafer elsewhere.

$$T_{read} = T\left(\frac{\varepsilon}{\hat{\varepsilon}}\right)^{1/4}, \qquad T = T_{SP}\left(\frac{\hat{\varepsilon}}{\varepsilon}\right)^{1/4} \tag{1.1}$$

At a setpoint of 1050 °C a film stack whose emissivity is 10 % higher than assumed is held at 1018.85 °C, 31.15 K low, while the trend shows the setpoint exactly [19]. Chapter 9 builds its pyrometers on this relation and finds the same 31.2 K, and its sheet resistance comes out 26.3 % high on a nitride cap that the pyrometers take for bare silicon. A real narrow-band pyrometer makes about a third of this error, which the chapter also computes.

A metrology offset does the same to a run-to-run loop. If M-1's after-etch microscope reads 0.5 nm high, a loop with integral action settles the reported linewidth on 40 nm and the true one on 39.5 nm, and Chapter 6 measures 0.8 nm of real material removed for an offset of 0.8 nm. A third kind of wrong measurement is one whose physical basis has gone. In Chapter 8 a spark halves the extraction voltage for milliseconds, the current reading stops describing the source, and a gate that marks it invalid cuts the beam overshoot after the spark from 14.75 to 0.29 %.

1.4.3 Run-to-Run Control

M-1's etcher trims the developed line, and the finished linewidth of lot k is $y_k = d_k + b\,u_k + v_k$, where u_k is the trim time, $b = -1.0\,\mathrm{nm/s}$ is the process gain, d_k is everything else the line does and v_k is noise. At the design point, a developed line of 60 nm, a fixed etch bias of 5 nm and 15 s of trim give 40 nm. Nothing measures d_k. A run-to-run controller estimates it from the readings and sets the next recipe so that the estimate lands on target [14], as in Equation (1.2).

$$\hat{d}_k = \lambda\left(y_k - \hat{b}\,u_k\right) + (1-\lambda)\,\hat{d}_{k-1}, \qquad u_{k+1} = \frac{y^* - \hat{d}_k}{\hat{b}} \tag{1.2}$$

The estimate is an exponentially weighted moving average of the disturbance each reading implies, and $0 < \lambda \le 1$ sets how much of each reading it believes. Substituting the recipe into the average gives the incremental form of Equation (1.3).

$$u_{k+1} = u_k - \frac{\lambda}{\hat{b}}\left(y_k - y^*\right) \tag{1.3}$$

It is a discrete integral controller with no proportional or derivative term, sampled once a run, and $\lambda/\hat{b}$ is its integral gain. With $b < 0$ the trim rises when the line is wide, so the loop is direct acting. Chapter 6 writes the same pair of equations for its etch trim and its exposure dose, Chapter 9 corrects its soak time from each wafer's sheet resistance, and Chapter 10 its plating time from each wafer's thickness. Chapter 5 runs a hundred of these loops, one per zone, on the measured linewidth map, and Chapter 11 one per profile mode.

> ✦ DID YOU KNOW? **A Third of the Cost Is Patterning**
>
> Lithography is the most frequently used process in a chip fab and typically accounts for 30 to 35 percent of the total process cost. A linewidth held within 10 % of its target is customarily required of a high-performance process [10].

The weight trades noise against drift [13]. With the gain known and the reading available before the next run, the error $e_k = y_k - y^*$ obeys $e_{k+1} = (1-\lambda)\, e_k$ plus the change in the disturbance between the two runs. Against white noise of variance σ^2 the loop has nothing to correct and passes a filtered copy of the noise into the recipe, so the output variance rises to $2\sigma^2/(2-\lambda)$. At $\lambda = 0.4$ that is 25 % more variance, 11.8 % more spread, than no control at all. Against a drift of δ per run the loop must move the recipe every run, and it can do so only from an error, so it settles with the offset of Equation (1.4).

$$e_\infty = \frac{\delta}{\lambda} \tag{1.4}$$

A drift of 0.05 nm per lot leaves 0.125 nm at $\lambda = 0.4$ and 0.25 nm at 0.2. Chapter 6 measures this residual against the same formula, 0.2500 nm at a drift of 0.075 nm per lot and $\lambda = 0.30$. Chapter 11 adds a slope term for its wearing pad, the double EWMA that removes a steady ramp's offset in theory; the chapter measures how far it reduces the offset on a wearing pad.

The best weight depends on how the disturbance moves. Suppose M-1's d_k is a random walk with steps of 0.2 nm per lot, plus white noise of 0.5 nm. With the variances q of the steps and r of the noise, the error variance under Equation (1.2) with a known gain is Equation (1.5).

$$\sigma_e^2 = \frac{(1-\lambda)^2 q + \lambda^2 r}{1-(1-\lambda)^2} + (m+1)\, q + r, \qquad \frac{\lambda_*^2}{1-\lambda_*} = \frac{q}{r} \tag{1.5}$$

where m is the metrology delay of Section 1.4.4, zero for now. The weight λ_* that minimizes it is the steady gain of a Kalman filter for this disturbance [20]. For M-1, $q/r = 0.16$ and $\lambda_* = 0.328$, which gives a spread of 0.610 nm. Figure 1.4b shows the trade. A weight of 0.1 lets the walk get ahead, 0.688 nm, a weight of 0.8 chases noise, 0.677 nm, and with no drift any weight is worse than none.

1.4.4 Metrology Delay

M-1's after-etch reading of lot k arrives after lot $k+2$ has been etched, so it first changes the recipe of lot $k+3$. A step in d on lot j therefore produces $m+1 = 3$ lots at full error before the recipe moves at all, whatever the weight. If the controller simply adds a fraction of the newest reading's error to its estimate, the closed loop is Equation (1.6), with $K = \lambda\xi$ the loop gain [21].

$$e_{k+1} = e_k - K\, e_{k-m}, \qquad z^{m+1} - z^m + K = 0 \tag{1.6}$$

Placing a root on the unit circle, $z = \mathrm{e}^{j\omega}$, separates the characteristic equation into a phase condition that fixes $\omega = \pi/(2m+1)$ and a magnitude condition that gives the stability boundary of Equation (1.7).

$$K_{max} = 2 \sin \frac{\pi}{2(2m+1)} \tag{1.7}$$

It returns 2, 1, 0.618, 0.445 and 0.347 for $m = 0$ to 4, so the first lot of queue halves the usable gain. The weight and the gain ratio enter only as a product. A controller that believes the trim rate is half the true one runs at twice the loop gain it was tuned for, so $\lambda = 0.4$ with $\xi = 2$ gives $K = 0.8$ and diverges at $m = 2$.

Figure 1.4a shows the cost on M-1 at $\lambda = 0.4$. With no queue the error falls by 0.6 per lot and stays within 10 % of the step after five lots. With $m = 2$ the dominant root is 0.882, the error rings with a period near ten lots, and it stays within 10 % only after 20 lots. At $\lambda = 0.7$, outside the boundary, it grows.

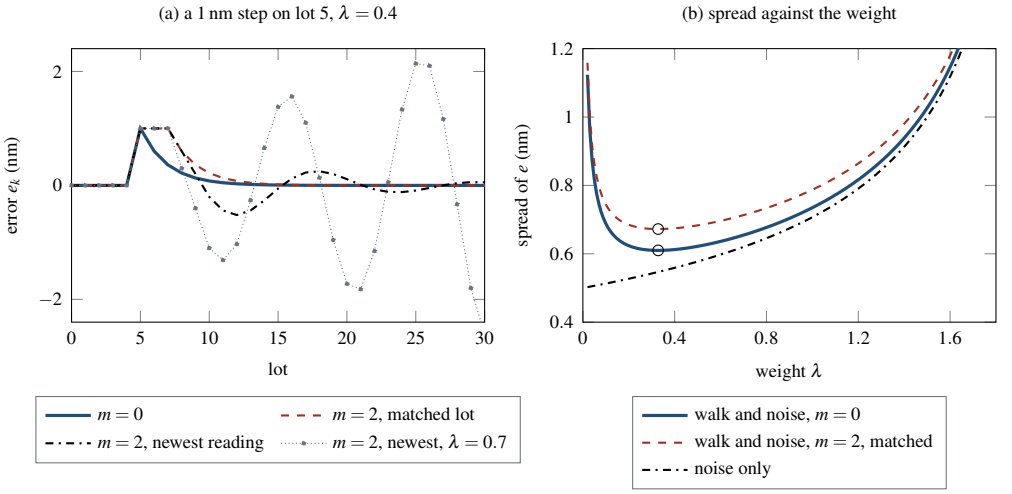


Figure 1.4 Run-to-Run Control of M-1's Trim. (a) The Error After a 1 nm Step With No Queue (Thick Solid), With a Two-Lot Queue and the Matched-Lot Estimator (Dashed), With the Newest-Reading Estimator (Dash-Dotted) and With That Estimator at a Weight Beyond the 0.618 Boundary (Dotted, Dots). (b) The Spread Against the Weight for a Random Walk of 0.2 nm per Lot Plus 0.5 nm of Noise, Without and With the Queue, and for Noise Alone. Open Circles Mark the Best Weight, 0.328.

The boundary belongs to the estimator, not to the plant. A controller that records which recipe made each lot can regress the delayed reading on that recipe, $\hat{d}_k = \lambda(y_{k-m} - \hat{b}\,u_{k-m}) + (1-\lambda)\hat{d}_{k-1}$, and with the gain known the delayed term then cancels. The error still sits at full value for $m+1$ lots, then falls by $1-\lambda$ per lot at any queue length, as the dashed curve shows: inside 10 % after seven lots. Chapter 6 measures both estimators on one line and finds the matched-lot boundary 3.24 times the other at $m = 2$. Two costs remain with the matched lot. The dead band of $m+1$ lots stays, and against a drifting disturbance each lot of queue adds the variance q of one more step, as the $(m+1)\,q$ term of Equation (1.5) states. For M-1 the queue raises the best spread from 0.610 to 0.672 nm without moving the best weight.

Sampling one lot in n_s lengthens the time between corrections by n_s with no change of λ, and Chapter 6 measures a second cost in the feedforward of Section 1.4.7. A reading that describes an old state of the tool must also be compared with the model of that time. Chapter 11 replays each late removal map against the pad state and recipe of the wafer that made it [22], and Chapter 10 compares each analysis with its model at the moment the sample left the bath. Its continuous version of the boundary is a trim gain of 210 for a 42 min analysis age, which would rise to 744 with an age of 2 min. Delay in continuous loops is the subject of Volume 5.

> ✦ DID YOU KNOW? **Most Products Run Only a Few Lots**
>
> A fab runs many products on each tool, and few of them often. In one manufacturing study more than 70 percent of the products had fewer than ten runs [10], so a controller that keeps one estimate per product can wait a long time between its readings.

1.4.5 Multizone Thermal Control

M-1's bake plate has five zones in a row. Each zone holds $C_z = 10\,\mathrm{J/K}$, loses $U = 0.5\,\mathrm{W/K}$ to the surroundings and exchanges $k_x = 1.0\,\mathrm{W/K}$ with each neighbor, so its temperature rise θ_i above ambient obeys Equation (1.8).

$$C_z \frac{\mathrm{d}\theta_i}{\mathrm{d}t} = P_i - U\theta_i - k_x \sum_{j \in \mathcal{N}(i)} (\theta_i - \theta_j), \qquad C_z \frac{\mathrm{d}\theta}{\mathrm{d}t} = \mathbf{P} - \mathbf{A}\theta \tag{1.8}$$

where $\mathcal{N}(i)$ is the set of neighbors and $\mathbf{A}$ is the conductance matrix. Its steady gain is not one number. A power pattern shaped like an eigenvector of $\mathbf{A}$ produces a temperature pattern of the same shape, and for zones in a row with insulated ends the eigenvectors are cosines with j half-waves across the plate. Their gains and time constants are Equation (1.9).

$$G_j = \frac{1}{U + 2k_x[1 - \cos(j\pi/N)]}, \qquad \tau_j = C_z G_j, \qquad j = 0, \ldots, N-1 \tag{1.9}$$

On M-1, with $N = 5$, the uniform pattern has a gain of 2.0 K/W and a time constant of 20 s, and the pattern that alternates from zone to zone has 0.243 K/W and 2.43 s, 8.24 times less, as Figure 1.5a shows. A smooth correction is cheap and a ragged one costs heater power, 0.5 W per kelvin against 4.12 W. The number of zones does not set the resolution a plate can deliver. The ratio of loss to coupling does [23].

A loop on one zone sees a third gain. With its neighbors held at their setpoints by their own loops, the center zone answers its heater with $1/(U + 2k_x) = 0.4\,\mathrm{K/W}$ and a time constant of 4 s. Tuned on that response, with the loop gain $K_c G = 1$ and the integral time equal to the time constant, $K_c = 2.5\,\mathrm{W/K}$ and $\tau_I = 4\,\mathrm{s}$. The tuning is right for a move of one zone, which reaches 90 % in 8.1 s without overshoot. When all five setpoints move together the loops push the uniform pattern, whose gain is five times larger and whose time constant is five times longer, so the loop gain is 5 and the integral time is a fifth of the plant's. The center zone overshoots 19.6 %, as the dash-dotted curve of Figure 1.5b shows. Tuned instead on the uniform pattern, $K_c = 0.5\,\mathrm{W/K}$ and $\tau_I = 20\,\mathrm{s}$, the loops move all five zones without overshoot in 46 s to 90 % but bring a single zone only to 56 % in a minute. Neither tuning is wrong for the step it was tested on. The relative gain of the center zone is 1.77, positive and of order one, so the one-to-one pairing stands, and it says nothing about the direction in which all loops move at once.

Chapter 5 meets exactly this on a plate of a hundred zones, where the uniform pattern has 4.43 times the single-loop gain and a tuning made on one zone overshoots 18.8 % when all zones move. Its remedy is a model of the coupling. The power that a setpoint profile needs in steady state is $\mathbf{A}\theta_{SP}$, one sparse product per scan with no inverse in it, and supplying it as feedforward leaves the loops only a trim. The plate also has a one-sided actuator. A heater cannot cool, so a zone that must come down waits for its own losses. Chapter 9 adds two features of lamp heating. Its lamps light each other's zones, so it decouples with the inverse of a five-by-five gain matrix [24]. Its loss is radiation, so the wafer's time constant falls from 13.7 to 2.4 s between 400 and 1050 °C [25, 26]. The bake plate matters because the resist integrates an Arrhenius rate, and Chapter 5 derives 3.2 nm of linewidth per kelvin of bake [27]. Chapter 11 has the same spatial structure in pressure. Eight carrier chambers act on 33 radial points through broad footprints, so a ragged radial target is out of reach rather than badly tuned.

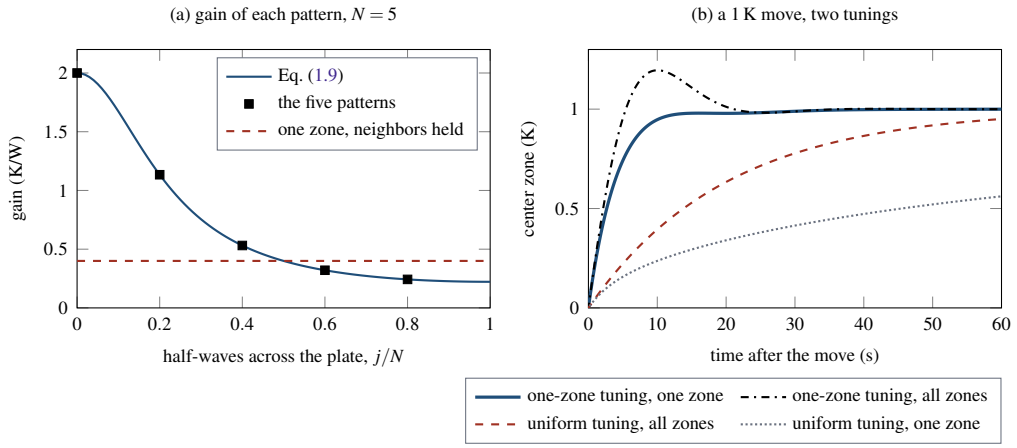


Figure 1.5 M-1's Five-Zone Bake Plate. (a) The Steady Gain of Each Spatial Pattern of Heater Power (Solid Curve, Squares at the Five Patterns) Against the Gain One Zone Sees With Its Neighbors Held (Dashed). (b) The Center Zone After a 1 K Setpoint Move of One Zone or of All Five, Under Loops Tuned on One Zone (Thick Solid, Dash-Dotted) or on the Uniform Pattern (Dashed, Dotted).

1.4.6 Endpoint Detection

M-1's etcher removes a film of $H = 150\,\text{nm}$ at a nominal rate of 3 nm/s. The rate varies by $\pm v = \pm 4\,\%$ across the wafer, so the film clears first where the rate is highest and last where it is lowest. Once a point has cleared, the etch attacks the layer underneath at a tenth of the rate. The main etch must run until the slowest point clears, and any time beyond that cuts into the underlayer where the film is already gone. The wafer passes if every point clears and the underlayer loses no more than 3 nm. From wafer to wafer the rate also moves, here between 0.94 and 1.06 of nominal as a freshly cleaned chamber seasons.

A timed recipe stops at a fixed $t_{end} = (1 + f_o)\,H/r_0$ after the nominal clearing time H/r_0. To clear the slowest wafer it needs an over-etch f_o of 10.8 %, and then the fastest wafer loses 3.32 nm of underlayer and fails. No timed recipe passes both ends. The rate band a timer can serve is $\pm 5.1\,\%$. An endpoint detector instead watches the light of an etch product, which falls as the film clears [7]. For a uniform spread of local rates between $r(1 - v)$ and $r(1 + v)$, half the wafer has cleared at H/r, and a detector that fires there and adds an over-etch f_o in proportion to its own firing time gives the conditions of Equation (1.10).

$$1 + f_o \geq \frac{1}{1 - v}, \qquad L = \frac{H}{S}\left(1 + f_o - \frac{1}{1 + v}\right) \tag{1.10}$$

where L is the underlayer loss at the fastest point and S the selectivity, 10 here. Neither condition contains the rate. The endpoint needs 4.17 % of over-etch to clear every wafer, and the non-uniformity alone sets it. M-1's detector filters the signal with a lag of 0.5 s and fires when about 62 % of the wafer has cleared. At 5 % over-etch every wafer from 0.90 to 1.10 of the nominal rate clears and loses 1.53 to 1.56 nm, as Figure 1.6 shows. The detector is not a feedback loop. It has no setpoint and no actuator, and it decides when the step ends.

Chapter 7 measures the same contrast on a chamber after a wet clean. Three of its four chamber loops hold their setpoints to four decimal places, the etch rate rises 6 % over eight wafers as the walls season, and a timed recipe scraps the first three. With the emission endpoint all eight pass, and the main etch shortens from 93.6 to 87.8 s as the walls season. In its graded cassette the endpoint is the single most valuable change. Other chapters end a step on an event in the same way. Chapter 2 harvests its rods at a diameter, Chapter 8 repaints the part of the wafer whose dose a spark interrupted,

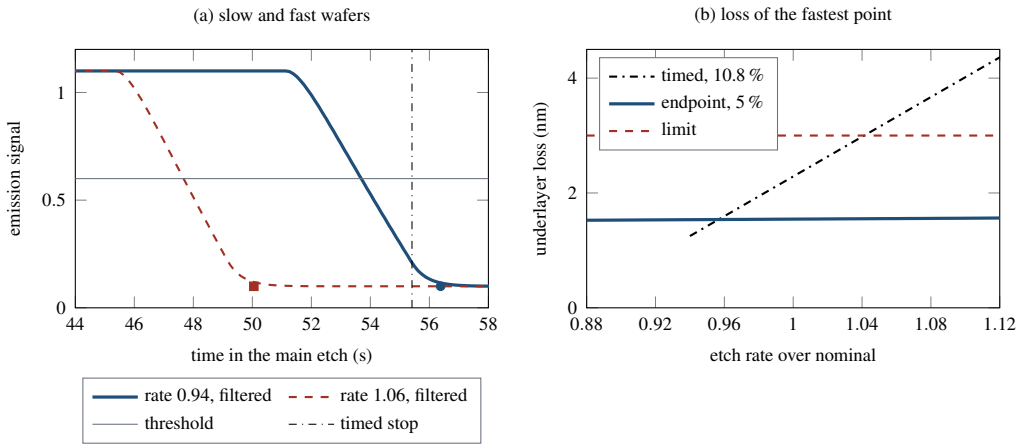


Figure 1.6 Ending M-1's Main Etch. (a) The Filtered Emission Signal of a Wafer Etched at 0.94 of the Nominal Rate (Solid) and at 1.06 (Dashed), the Detector's Threshold (Thin) and the Stop of a Timed Recipe (Dash-Dotted). The Dot and the Square Mark Where the Endpoint Recipe Stops Each Wafer. (b) Underlayer Loss Against the Etch Rate. The Timed Recipe (Dash-Dotted) Leaves Wafers Below 0.94 Uncleared and Exceeds the Limit (Dashed) Above About 1.04. The Endpoint Recipe (Solid) Clears Every Wafer With About 1.5 nm of Loss.

and Chapter 10 calls a dump of the bath when bleeding can no longer reopen its window. Batch endpoints in general are the subject of Volume 8.

1.4.7 Feedforward Between Process Steps

The developed linewidth of an M-1 lot is measured before the lot is etched, so the etcher can cancel the error that lithography left. Suppose lithography varies the developed line from lot to lot with a spread $\sigma_L = 1.0\,\text{nm}$, the microscope reads it with a spread $\sigma_M = 0.4\,\text{nm}$, and the etch adds $\sigma_E = 0.3\,\text{nm}$ of its own. Adding g times the measured deviation, divided by the trim rate, to the trim of the same lot leaves the variance of Equation (1.11).

$$\sigma_y^2 = (1-g)^2\sigma_L^2 + g^2\sigma_M^2 + \sigma_E^2, \qquad g_* = \frac{\sigma_L^2}{\sigma_L^2 + \sigma_M^2} \tag{1.11}$$

Without feedforward the finished spread is 1.044 nm. With $g = 1$ the whole lithography term leaves the expression and is replaced by the microscope's, 0.500 nm, and the best gain $g_* = 0.862$ trims it to 0.477 nm. Full feedforward helps exactly when the measurement is quieter than the process it measures. It acts only on lots that were measured, so measuring one lot in three keeps a third of the benefit, 0.900 nm. Feedback cannot match it. A white lot-to-lot error is gone before any later reading can describe it, and feedback at any weight adds spread to it, as Section 1.4.3 showed.

Chapter 6 implements this path. At an unchanged feedback weight it measures the spread fall from 1.852 to 0.850 nm, below the 1.476 nm of no run-to-run control at all. It also shows that the feedback estimator must regress on the feedback part of the recipe only, or it will undo what the feedforward did. Chapter 11 feeds the measured incoming film profile to its recipe, and Chapter 10 doses its additives in proportion to the charge passed, a feedforward from the load. Two more chapters feed forward within a step. The scanner of Chapter 4 reads the wafer surface ahead of the slit and commands the stage from it [28], with a lead distance equal to the scan speed times the stage's lag, so the right distance moves with the throughput. The bake plate of Chapter 5 adds a power pulse when a cold wafer lands, a disturbance whose time and size are known in advance, and holds the zone integrators while the pulse runs so that feedback does not undo it. Feedforward in general is the subject of Volume 3.

Table 1.1 For Each Chapter, What the Tool Measures While It Runs, What Is Measured on the Product and When, and How the Next Wafer or Lot Is Corrected

Ch.	In the tool, fast	On the product, and when	Next wafer or lot
2	Rod surface pyrometer, supply volts and amperes	None in the batch, center temperature inferred from the supply	Within the batch: gain schedule on diameter, override on inferred center
3	Camera diameter, heater power	No instrument reports the pull rate limit	Within the batch: diameter on pull rate, heater mid-ranges the pull rate
4	Surface ahead of the slit, dose in the slit	Overlay from sampled fields of the wafer	Preview feedforward with a trailing trim, grid fit applied open loop
5	A hundred zone thermocouples, 0.1 s	Linewidth map, every wafer	A run-to-run offset per zone, a pulse at each landing
6	Plate temperature	Developed line at once, etched line two lots late	Dose and trim by EWMA, trim fed forward from the developed line
7	Pressure, density, bias, wafer temperature, emission	Seven criteria on each finished wafer	Main etch ends on the emission endpoint
8	Extraction current every 1 ms, beam current	Dose from the scan ledger	Gated cascade, local repaint of missed dose
9	Five pyrometers, 10 ms	Sheet resistance after each wafer	Soak time corrected wafer to wafer
10	Charge passed, temperature	Additives every 2 h, 42 min old, thickness each wafer	Dosing by charge with a sampled trim, plating time run to run
11	Chamber pressures, pad counters	Removal map one or two lots late, incoming map	Constrained recipe, age-aware EWMA per mode, fresh-pad estimate

1.4.8 What Each Tool Measures and How the Next Wafer Is Corrected

A fab runs these techniques in layers: fast loops on tool sensors inside the step, an endpoint or a supervisor that ends the step, and run-to-run and feedforward corrections between steps on metrology results [29, 30]. Table 1.1 lists, for each chapter, what the tool measures while it runs, what is measured on the product and when, and how the next wafer, lot or run is corrected. Two chapters have no product measurement between runs because the run is one batch. Chapter 2 schedules its gain on the diameter of a rod that grows a hundredfold in area, and Chapter 3 returns its fast, limited pull rate to mid-range with the slow heater. Chapter 12 adds the measured results and compares what each tool measures, when, and how the next wafer or lot is corrected.

1.5 Instruments and Final Elements in This Volume

Every loop of this volume begins at a sensor and ends at a final element, and in a fab the two sit at two levels. Figure 1.7 lists what the controllers of each chapter read and what they move. Inside the tools a temperature is read by a thermocouple or a probe in four chapters and by a radiation pyrometer in two, and a current, a voltage or a charge serves four. Heater, lamp or rod power is moved in five chapters, and a supply that feeds a plasma, an ion source or a laser in three. Five chapters read a measurement made on the wafer after the step, and in each of them a final element is a number in the recipe of a later wafer or lot.

The temperature instruments differ in what they touch. A thermocouple is two wires of dissimilar metals joined at one end, which give a voltage roughly proportional to the temperature of the junction. It is less accurate than a resistance thermometer but rugged, and its most common failure is to burn out open [31]. Each of the hundred zones of Chapter 5 reads one, as drawn in Figure 1.8. A pyrometer, drawn in Figure 1.9, concentrates the radiation of a surface onto a small sensing element and touches nothing. It reveals only the surface, and the only certain way to know an object's emittance is to test its radiation at a known temperature [31]. That is the error of Section 1.4.2, met in Chapter 2 and Chapter 9.

The product is read elsewhere. A linewidth, a thickness or a sheet resistance is measured on a metrology tool that serves many process tools, and the correction it causes is written into a later

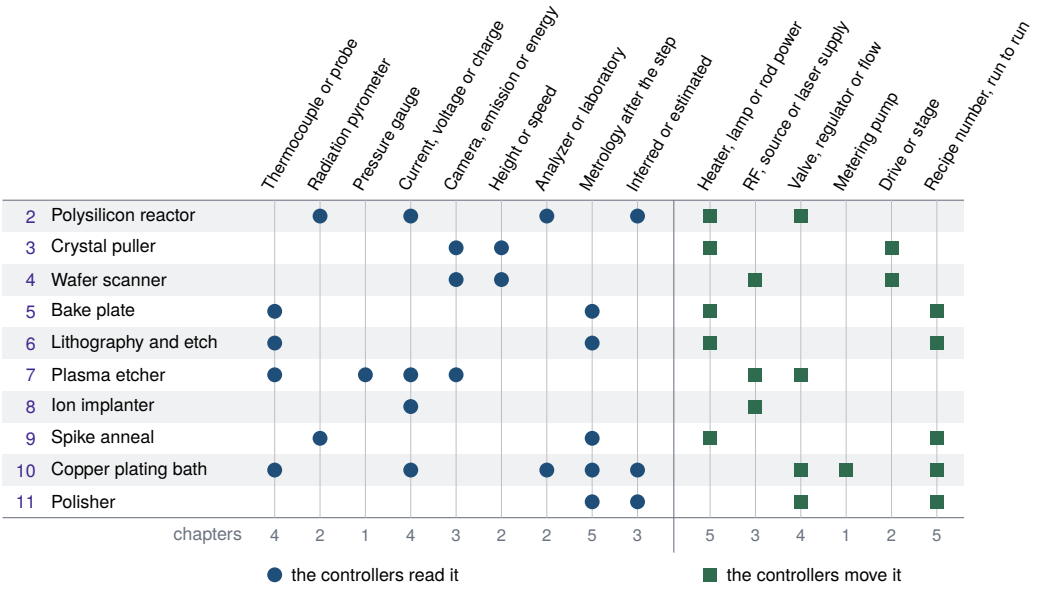


Figure 1.7 Sensors and final elements by chapter. A round mark means that the chapter's controllers read the device, a square mark that they move it. A mark is set only where the chapter's variables table, its instrument text or its section on technology in practice names the device or its reading, so a blank says that the model leaves the device out and not that a real tool lacks it. Notice that the five chapters that read metrology after the step are the five whose final element is a recipe number.

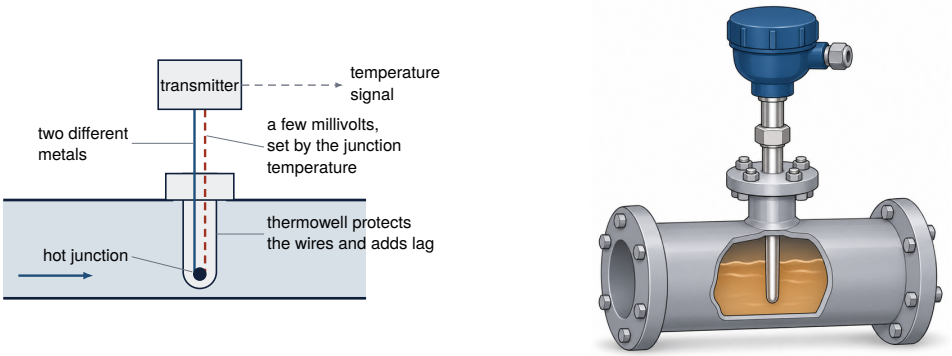


Figure 1.8 Principle of a thermocouple in a thermowell, with a view of the device as installed. Notice that the reading is the temperature of the junction, so a junction that sits in a plate reads the plate and not the wafer on it.

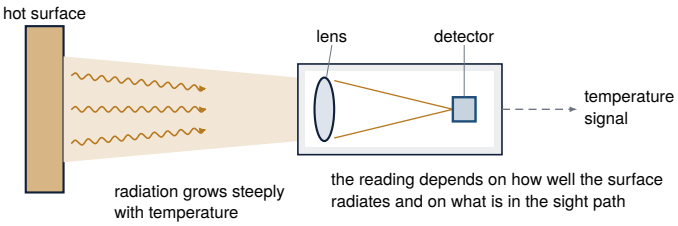


Figure 1.9 Principle of a radiation pyrometer. Notice that the instrument converts radiance to temperature with an emissivity it is given, so a film that changes the true emissivity moves the wafer while the trend still shows the setpoint.

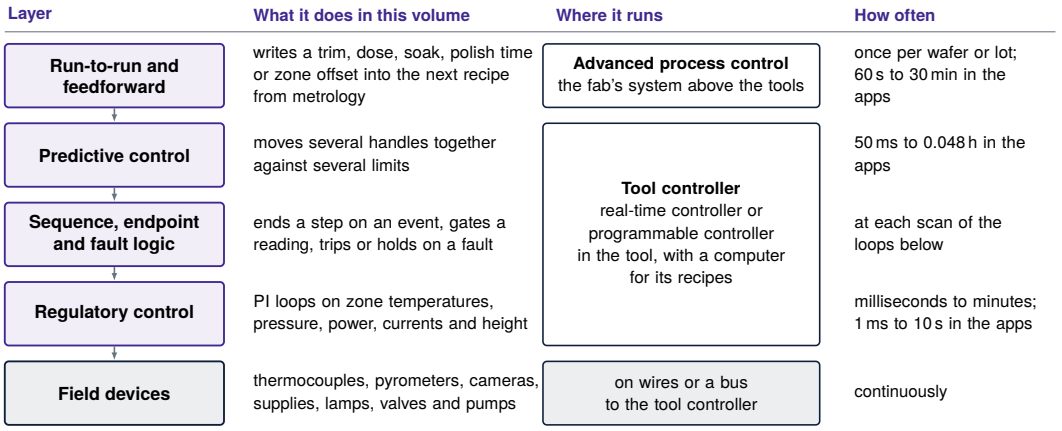


Figure 1.10 The control layers of this volume, the platform that hosts each and how often each acts. The range of loop periods is that of [32], the advanced process control system that of [10], and the ranges in the apps those of the chapters. Notice that the recipe layer acts once per wafer or lot, while the loops below it act many times within one step. In the polysilicon reactor, a plant rather than a fab tool, the predictive controller sits at the supervisory level above the loops (Section 2.5.4).

Table 1.2 What Every App Provides. Standalone Apps Are Described in Their Own Chapters.

Element	What it does
Run, Pause, Reset	Start, hold and restart the run. Reset keeps the reader's settings.
Speed	Simulated time per second of real time. It changes the pace and never the result.
Manual and Auto	Hand each actuator to the reader or to the controller, without a bump.
Scenario list	Numbered cases, each with a briefing, a completion rule and a reload button.
Control panel	Setpoints, manual outputs, tuning, options and disturbances, at a Beginner or an Advanced level.
Trends	SP and PV above the MV for each loop, with limits, event marks and a cursor.
Scorecard	IAE, overshoot, settling time, MV travel and time at a limit for each loop.
Learn card	Predict, Try It, two hints and Explain, opened by the reader in that order.
Show Model	Equations, live variable values, editable parameters and assumptions.
Export CSV	Every logged sample of the run as a text file with a header row.
Instructions	A separate page with background, an operating guide and the model's limitations.

recipe. Run-to-run control of this kind adjusts the recipe lot by lot or wafer by wafer, and most fabs combine it with fault detection and classification in an advanced process control system [10].

Figure 1.10 places the layers of the volume on their platforms. Control loops in operational technology run with cycle times from milliseconds to minutes, many on real-time operating systems [32]. The apps run their regulatory loops every 1 ms in Chapter 4 up to every 10 s in Chapter 10, their predictive controllers every 50 ms in Chapter 9 up to every 0.048 h in Chapter 2, and their recipe corrections once per wafer or lot, from every 60 s in Chapter 9 to every 30 min in Chapter 6.

1.6 The Simulator Framework

Every chapter of this book has an app that runs in a web browser, collected with the others at https://apmonitor.com/apps/. About 240 apps exist across the series. All but a few are built on one shared framework, so the engine, the PID algorithm, the trends and the data export behave the same way in every chapter. The exceptions are standalone studios, such as the cascade studio and the two Temperature Control Lab apps, which follow the same ideas in their own code. This section describes the framework once and states where the standalone apps differ. Table 1.2 lists what a reader meets.

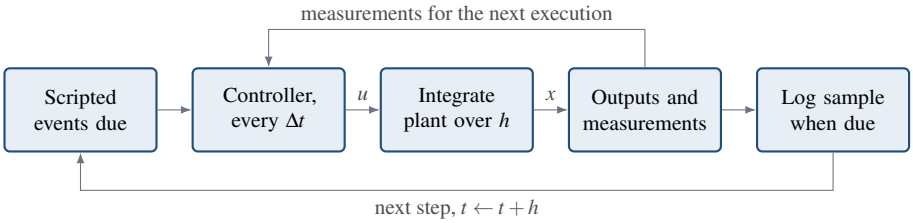


Figure 1.11 Work of the Engine in One Integration Step

1.6.1 The Simulation Engine

Each app carries a plant model written as ordinary differential equations $\frac{dx}{dt} = f(x,u,d,p)$ in the states x, the inputs u, the disturbances d and the parameters p. The engine integrates the model with the classical fourth-order Runge–Kutta method on a fixed step h set for the fastest dynamics of the app. Some apps supply their own step, such as explicit Euler for a discontinuous right-hand side or an exact update for a model driven by events.

Three clocks run at separate rates. The plant advances every step h. The controller executes every control period Δt, equal to h unless the app declares a slower controller, and holds its outputs between executions. The log takes one sample at a fixed interval of simulated time for the trends, the scorecard and the exported file. The order of work in one step is shown in Figure 1.11. The controller also executes once when a scenario is loaded, so every run starts from an initialized controller.

Drawing the screen is kept apart from these clocks. A speed of 5 means five units of the app's time base per second of real time. Speed, frame rate and pauses change how fast a run is shown and never what it computes, so a phone and a workstation produce the same trajectory and the same file.

Noise comes from a deterministic pseudorandom generator. Its seed is declared by the app, or by a scenario that needs its own noise record, and it is restored whenever a scenario is loaded or Reset is pressed. The framework gives the reader no seed setting, so *default seed* in this book means the seed the app ships with. A run repeats bit for bit when the same scenario is loaded and the same settings are made before Run. A change made by hand during a run lands on a different step each time, which is one reason numeric answers carry a tolerance. The standalone simulators draw noise from the browser's own generator, so a noisy run there repeats in its statistics and not sample for sample.

Run and Pause start and hold the clock. Reset restarts the run at $t = 0$, paused, and keeps the reader's settings: every setpoint, tuning value, option, model parameter, disturbance slider and mode the reader changed is put back before the plant and the controller are initialized. Scripted events replay, and a one-shot disturbance button pressed during the run is cleared. The reload button beside the scenario list restores the scenario's own settings instead.

1.6.2 The Velocity-Form PID Algorithm

One PID implementation serves the PID loops of every framework app. It is the velocity (incremental) form: each execution adds the change in the proportional and derivative terms to a running sum, and the integral is a separate term, as shown in Equations (1.12) to (1.15) for execution k and control period Δt.

$$e_k = a\,(\mathrm{SP}_k - \mathrm{PV}_k), \qquad P_k = K_c\,e_k, \qquad D_k = -a\,K_c\,\tau_D\,\frac{\mathrm{PV}^f_k - \mathrm{PV}^f_{k-1}}{\Delta t} \tag{1.12}$$

$$S_k = S_{k-1} + (P_k - P_{k-1}) + (D_k - D_{k-1}) \tag{1.13}$$

$$v_k = u_b + S_k + I_{k-1}, \qquad u_k = \mathrm{sat}(v_k) \tag{1.14}$$

$$I_k = I_{k-1} + \left[\frac{K_c}{\tau_I}\,e_k + \frac{u_k - v_k}{T_t}\right]\Delta t \tag{1.15}$$

where $a = \pm 1$ is the controller action, PV^f is the measurement after an optional first-order filter, u_b is the output bias, S is the running sum of proportional and derivative increments, v is the output before limiting, u is the output sent to the actuator, I is the integral term and T_t is the tracking time constant, equal to τ_I unless a loop sets another value. The function sat clamps v_k to the actuator range and then applies a rate limit if the loop declares one.

The limit acts on the emitted output u_k only. The sum S_k is never clamped, so it equals $P_k + D_k$ and the output before limiting is the position form $u_b + K_c e_k + D_k + I_{k-1}$. Anti-windup touches the integral alone. While the output is limited, the last term of Equation (1.15) drives I toward the value that reproduces the output applied, which is back-calculation [33]. With no integral action, Equation (1.15) is skipped. Clamping the running sum would instead discard each increment that crosses a limit and keep each one that returns, so symmetric noise would push the output off the limit even in a proportional-only loop.

The derivative acts on the measurement, so a setpoint step gives a proportional kick and no derivative kick. A change of K_c acts on the whole error at the next execution, as in the position form. The action $a = +1$ suits a process whose PV rises with the MV, a reverse-acting controller in the ISA convention, and $a = -1$ gives the direct-acting one.

In Manual the controller keeps executing and sets $I_k = u_{\mathrm{man}} - u_b - S_k$, so the output it would send equals the manual value u_{man}. The first automatic output differs from it by one period's change in P and D. In Auto the engine copies each controller output to the manual station, so a switch to Manual is also bumpless. The standalone studios with PID loops use velocity-form code of their own, with the limit on the output and anti-windup on the integral alone, in which a gain change applies only to later increments.

1.6.3 The Predictive Control and Estimator Options

More than 100 apps offer model predictive control (MPC) as an option beside the structure they open with, which stays the default. Volume 9 develops the theory. In most of them the controller plans with a small linear model in discrete time as shown in Equation (1.16).

$$x_{k+1} = A\,x_k + B\,u_k + E\,w_k + c, \qquad y_k = C\,x_k \tag{1.16}$$

where w is a measured or scheduled disturbance and c is a constant left by linearization. The model comes from a step test at the design point, from the app's balances linearized there or again at every decision, or from the model the reader fits. It never reads the true states or parameters of the run, so a parameter the reader changes in Show Model is model mismatch, as on a real plant.

At each decision the controller minimizes the objective of Equation (1.17) over a horizon of N steps.

$$J = \sum_{k=1}^{N} (y_k - r_k)^{\top} Q\,(y_k - r_k) + \sum_{j} \Delta u_j^{\top} R\,\Delta u_j + \sum_{i} \rho_i\, s_i^2 \tag{1.17}$$

where r is the setpoint, Q weights the tracking error, R is the move suppression on each input change Δu_j and ρ_i is the price of a slack $s_i \geq 0$ that lets an output band be broken when it cannot be held. Actuator limits are hard constraints. The quadratic program is solved in the browser by a dual active-set method [34]. The first move is applied, and the plan is made again at the next decision.

The plan starts from an estimate of the model state. About half of these apps use a steady-state Kalman filter on the model with an added disturbance state, which absorbs what the model did not predict and removes steady offset without an integral term [35]. Others correct the model output with a filtered bias to the same end. About a dozen apps also offer moving horizon estimation (MHE), which solves the estimation problem over a window of recent samples and can hold bounds that a Kalman filter cannot, such as a concentration that is never negative [36].

With the MPC selected, the tuning fields of the loops it replaces leave the Control panel. In most of these apps the Advanced level shows up to three settings, each a multiple of the shipped value: the prediction horizon, the move suppression and the band price. The trend draws the accepted plan

ahead of the present time, and a list of constraint prices shows which limits bind. A plan that is not optimal, not finite or outside its limits is rejected, and a fallback, in most apps the shipped controller, holds the actuator until a plan succeeds. The Model Fitting and Tuning Studio builds the same kind of controller from the reader's own data in its tab, MPC from the fit.

1.6.4 Scenarios, Progress and Data

An app teaches through scenarios, the numbered cases at the top of the Control panel, nine to thirteen in most apps. Loading one sets the model parameters, the controller settings, the disturbances, the mode, the initial state and any scripted events. The information button beside the list opens a briefing in four parts: what the case is, why it matters, what to do and one detail for the practicing engineer. The Learn card holds the lesson back, with Predict and Try It before the run, hints on request and Explain afterward.

Each scenario has a completion rule that the app judges from the run on the simulated clock: a run length, the scripted events passed, a setting made by the reader, a state held for a stated time or a target met. Choosing a case earns nothing. A completed case is marked in the list, and a badge row shows the count and a level, Explorer after one case, Practitioner at half and Completed at all. Progress is kept in the reader's browser with no account. A certificate of progress or of completion can be issued from it, as a record of self-paced work and not an assessed credential.

One case, usually the last, is the graded challenge. It states a target, such as an error band to hold through a sequence of upsets. Changing the model parameters or the disturbances voids the attempt in most apps, and a record made with Reveal suggested tuning says so.

Show Model opens the equations, live values of the variables, the editable parameters and the assumptions. The trends plot SP and PV above the MV for each loop, with actuator limits and marks at setpoint moves, disturbances, mode changes and scripted events. The scorecard reports for each loop the integral of the absolute error (IAE), overshoot, settling time, MV travel and share of time at a limit, with the app's own measures, and it does not combine them into one score. IAE and travel accumulate from Reset, so two runs are compared at the same clock reading.

Export CSV saves the log as a text file with a header row of column names that carry their units and one row per logged sample. In most apps the time column is in seconds, minutes or hours according to the run length. The log holds the most recent 60 000 samples unless the app sets another size. Apps with the MPC option begin the file with comment lines that start with #, which a script must skip.

An app exercise in this book names the scenario by its number and name as the app shows them, every setting that differs from the scenario's defaults, the run length in simulated time, the seed (the default seed unless stated), what to read and where (a trend, a scorecard tile or a column of the exported file) and a tolerance on a numeric answer. Settings are made before Run unless the exercise gives a time.

1.7 Working With AI Agents While Studying

A reader of this book is likely to have an AI assistant within reach, and the apps make its work easy to test. This section gives the editor's view of how to use one while working through a chapter. In short, the assistant may do the algebra and the scripting, and the reader keeps the checking.

What to delegate. Tasks with a clear input and a result that can be tested are good candidates. An assistant can derive the linearization of a chapter's balance at an operating point and the transfer function that follows from it. It can write a short script that reads the exported CSV file and fits a first-order-plus-dead-time model to a step test. It can propose a first tuning from that fit with a named rule, convert units, explain a term in a model or outline a report. The assistant knows nothing about the app until it is told. Give it the equations and parameter values from Show Model, the column names of the exported file and the scenario that produced the data, as described in Section 1.6.4. Ask it to state its assumptions and the units of every result.

Table 1.3 Roadmap of the Volume

Ch.	Plant	The lesson	Level	Time
2	Siemens polysilicon reactor	A gain that falls a hundredfold is scheduled on its cause, the diameter.	introductory	3 h
3	Czochralski crystal puller	A fast, limited actuator is returned to mid-range by a slow one.	intermediate	3 h
4	Wafer scanner	Where nothing can be measured, the command leads the process.	intermediate	2.5 h
5	Hundred-zone bake plate	Zone loops tuned one at a time overshoot together.	intermediate	2.5 h
6	Lithography and etch line	The queue, not the plant, limits the recipe loop.	introductory	2.5 h
7	Plasma etch chamber	The recipe and the endpoint matter more than the loops.	introductory	2.5 h
8	Ion implanter source	A measurement can stop meaning anything, and the loop must know.	advanced	2.5 h
9	Spike anneal chamber	A perfect trend can sit on the wrong wafer temperature.	advanced	2.5 h
10	Copper plating bath	A two-hour analysis limits the trim, so the dose follows the charge.	intermediate	2 h
11	Chemical mechanical polisher	A pad change decides what the estimator may still believe.	advanced	2.5 h

What to check, and how. An assistant returns a fluent answer whether or not the answer is right, so each result needs a test that it could fail. Four tests cover most of the errors a chapter will produce.

- *Units.* Every term of an equation carries the same units, and a process gain is in PV units per MV unit. A gain computed with the valve as a fraction is 100 times the gain of a loop whose valve is in percent.
- *Signs.* Decide from the physics which way the PV moves when the MV rises. Compare that with the sign of the fitted gain and with the controller action of Section 1.6.2.
- *Steady state.* Set the derivatives to zero. The model must return the operating point that Show Model reports, and a linearization must give zero deviation there.
- *A run on the app.* Load the named scenario at the default seed, enter the proposed numbers and run. The trend and the scorecard decide. A tuning that reads well on paper and holds the valve at a limit for half the run has failed.

A fitted model gets one more test: compare it with data it was not fitted to, such as a second step in the opposite direction. When a check fails, report the failure to the assistant with the evidence and ask for the cause, then check the correction the same way.

How the exercises train the habit. One exercise in each chapter is marked *Review the agent*. It presents a tuning, a model or a claim as an AI assistant might propose it, with a flaw, and asks the reader to find the flaw in about five minutes. The flaws are ordinary ones: a unit, a sign, a steady state that does not close, a result measured on the wrong scenario or at a different run length, or a limit that was ignored. Each is found with one of the tests above. The purpose is not distrust. An engineer who can say why an answer is right can hand an assistant much larger tasks and still stand behind the result. That larger picture, in which agents do most of the routine work and the engineer supervises and checks, is the subject of Chapter 12.

1.8 Chapter Roadmap

Table 1.3 lists the chapters with the lesson each adds to this one. The level says how much of Section 1.4 a chapter assumes. The time is an estimate for reading the chapter once and running each case of its app once, at about eight minutes a case.

Chapters 2, 6 and 7 form a first course: a schedule on a growing rod, a recipe loop through a metrology queue, and a step that ends on an endpoint. They rest on Sections 1.4.3, 1.4.4 and 1.4.6. Chapters 3, 4, 5 and 10 add mid-ranging, preview feedforward, spatial modes and a sparse late analysis. Chapters 8, 9 and 11 combine a measurement that can mislead with fast dynamics or a consumable that is replaced, and are best left until last. A reader who follows the wafer can read the chapters in their order, which is the order of Figure 1.2.

KEY POINTS

- Most steps of chipmaking leave nothing that can be measured while they run, so the property that matters is measured afterward, late and on a sample.
- A run-to-run controller is a discrete integral controller sampled once a run, and its weight trades noise against drift.
- A metrology queue halves the usable gain at its first lot, but the boundary belongs to the estimator, not to the plant.
- Zone loops tuned one at a time can overshoot when all zones move together.
- An endpoint needs an over-etch that the non-uniformity alone sets, while a timer must also cover the spread of the etch rate.

Where else this applies. Run-to-run and batch control in general return in Volume 8, dead time and estimation in Volume 5, feedforward in Volume 3 and interaction in Volume 6.

References

[1] J. Czochralski, "Ein neues Verfahren zur Messung der Kristallisationsgeschwindigkeit der Metalle," *Zeitschrift für Physikalische Chemie*, vol. 92U, no. 1, pp. 219–221, 1918.

[2] K. S. Gokhale and B. M. Moudgil, "Particle technology in chemical mechanical planarization," *KONA Powder and Particle Journal*, vol. 25, pp. 88–96, 2007.

[3] H. Schweickert, K. Reuschel, and H. Gutsche, "Production of high-purity semiconductor materials for electrical purposes," U.S. Patent 3,011,877, filed 11 June 1957 (priority Germany, 25 June 1956), 1961, granted 5 December 1961.

[4] J. H. Freeman, "A new ion source for electromagnetic isotope separators," *Nuclear Instruments and Methods*, vol. 22, pp. 306–316, 1963.

[5] M. Steinbuch, T. Oomen, and H. Vermeulen, "Motion control, mechatronics design, and Moore's law," *IEEJ Journal of Industry Applications*, vol. 11, no. 2, pp. 245–255, 2022.

[6] T. Kimoto, A. Wachi, S. Sakurai, and M. Mikami, "Automatic diameter control of Czochralski single crystal rods," *Transactions of the Society of Instrument and Control Engineers*, vol. 9, no. 5, pp. 595–600, 1973, in Japanese, with an English abstract.

[7] K. Hirobe and T. Tsuchimoto, "End point detection in plasma etching by optical emission spectroscopy," *Journal of The Electrochemical Society*, vol. 127, no. 1, pp. 234–235, 1980.

[8] M. Uehara, T. Sudo, and F. Kanatani, "Horizontal position detecting device," U.S. Patent 4,558,949, filed 17 September 1982, 1985.

[9] R. Singh, "Rapid isothermal processing," *Journal of Applied Physics*, vol. 63, no. 8, pp. R59–R114, 1988.

[10] A. V. Prabhu, "Performance monitoring of run-to-run control systems used in semiconductor manufacturing," Ph.D. dissertation, The University of Texas at Austin, 2008. [Online]. Available: http://hdl.handle.net/2152/17721

[11] D. Edelstein, J. Heidenreich, R. Goldblatt, W. Cote, C. Uzoh, N. Lustig, P. Roper, T. McDevitt, W. Motsiff, A. Simon, J. Dukovic, R. Wachnik, H. Rathore, R. Schulz, L. Su, S. Luce, and J. Slattery, "Full copper wiring in a sub-0.25 µm CMOS ULSI technology," in *International Electron Devices Meeting. IEDM Technical Digest*, Washington, DC, USA, 1997, pp. 773–776.

[12] D. J. Chivers, "Freeman ion source: An overview (invited)," *Review of Scientific Instruments*, vol. 63, no. 4, pp. 2501–2506, 1992.

[13] A. Ingolfsson and E. Sachs, "Stability and sensitivity of an EWMA controller," *Journal of Quality Technology*, vol. 25, no. 4, pp. 271–287, 1993.

[14] E. Sachs, A. Hu, and A. Ingolfsson, "Run by run process control: Combining SPC and feedback control," *IEEE Transactions on Semiconductor Manufacturing*, vol. 8, no. 1, pp. 26–43, 1995.

[15] H. Ito, "Chemical amplification resists: Laboratory curiosity to paradigm," *Journal of Photopolymer Science and Technology*, vol. 20, no. 3, pp. 319–332, 2007.

[16] G. S. May and C. J. Spanos, *Fundamentals of Semiconductor Manufacturing and Process Control*. Hoboken, NJ: Wiley, 2006.

[17] S. J. Qin, G. Cherry, R. Good, J. Wang, and C. A. Harrison, "Semiconductor manufacturing process control and monitoring: A fab-wide framework," *Journal of Process Control*, vol. 16, no. 3, pp. 179–191, 2006.

[18] A. A. Khan, J. R. Moyne, and D. M. Tilbury, "An approach for factory-wide control utilizing virtual metrology," *IEEE Transactions on Semiconductor Manufacturing*, vol. 20, no. 4, pp. 364–375, 2007.

[19] P. J. Timans, "Emissivity of silicon at elevated temperatures," *Journal of Applied Physics*, vol. 74, no. 10, pp. 6353–6364, 1993.

[20] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[21] R. P. Good and S. J. Qin, "On the stability of MIMO EWMA run-to-run controllers with metrology delay," *IEEE Transactions on Semiconductor Manufacturing*, vol. 19, no. 1, pp. 78–86, 2006.

[22] A. Chen and R.-S. Guo, "Age-based double EWMA controller and its application to CMP processes," *IEEE Transactions on Semiconductor Manufacturing*, vol. 14, no. 1, pp. 11–19, 2001.

[23] B. Bamieh, F. Paganini, and M. A. Dahleh, "Distributed control of spatially invariant systems," *IEEE Transactions on Automatic Control*, vol. 47, no. 7, pp. 1091–1107, 2002.

[24] C. D. Schaper, T. Kailath, and Y. J. Lee, "Decentralized control of wafer temperature for multizone rapid thermal processing systems," *IEEE Transactions on Semiconductor Manufacturing*, vol. 12, no. 2, pp. 193–199, 1999.

[25] C. D. Schaper, M. M. Moslehi, K. C. Saraswat, and T. Kailath, "Modeling, identification, and control of rapid thermal processing systems," *Journal of The Electrochemical Society*, vol. 141, no. 11, pp. 3200–3209, 1994.

[26] F. Roozeboom and N. Parekh, "Rapid thermal processing systems: A review with emphasis on temperature control," *Journal of Vacuum Science & Technology B*, vol. 8, no. 6, pp. 1249–1259, 1990.

[27] J. P. Cain, P. Naulleau, and C. J. Spanos, "Critical dimension sensitivity to post-exposure bake temperature variations in EUV photoresists," in *Emerging Lithographic Technologies IX*, ser. Proc. SPIE, vol. 5751, 2005, p. 1092.

[28] M. Tomizuka, "Optimal continuous finite preview problem," *IEEE Transactions on Automatic Control*, vol. 20, no. 3, pp. 362–365, 1975.

[29] T. F. Edgar, S. W. Butler, W. J. Campbell, C. Pfeiffer, C. Bode, S. B. Hwang, K. S. Balakrishnan, and J. Hahn, "Automatic control in microelectronics manufacturing: Practices, challenges, and possibilities," *Automatica*, vol. 36, no. 11, pp. 1567–1603, 2000.

[30] E. Del Castillo and A. M. Hurwitz, "Run-to-run process control: Literature review and extensions," *Journal of Quality Technology*, vol. 29, no. 2, pp. 184–196, 1997.

[31] T. R. Kuphaldt, *Lessons in Industrial Instrumentation*, version 2.33 ed. Open textbook, Creative Commons Attribution 4.0 license, ibiblio.org/kuphaldt/socratic/sinst, 2024.

[32] K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, A. Hahn, S. Saravia, A. Sherule, and M. Thompson, "Guide to operational technology (OT) security," National Institute of Standards and Technology, Gaithersburg, MD, Special Publication 800-82r3, 2023.

[33] K. J. Åström and L. Rundqwist, "Integrator windup and how to avoid it," in *Proc. 1989 American Control Conference*, 1989, pp. 1693–1698.

[34] D. Goldfarb and A. Idnani, "A numerically stable dual method for solving strictly convex quadratic programs," *Mathematical Programming*, vol. 27, no. 1, pp. 1–33, 1983.

[35] G. Pannocchia and J. B. Rawlings, "Disturbance models for offset-free model-predictive control," *AIChE Journal*, vol. 49, no. 2, pp. 426–437, 2003.

[36] C. V. Rao, J. B. Rawlings, and D. Q. Mayne, "Constrained state estimation for nonlinear discrete-time systems: Stability and moving horizon approximations," *IEEE Transactions on Automatic Control*, vol. 48, no. 2, pp. 246–258, 2003.

1.9 Exercises

1.1 Theory and calculation · 4 min

A run-to-run controller of the form of Section 1.4.3 runs at $\lambda = 0.25$ with the process gain known and no metrology delay.

(a) The only disturbance is white lot-to-lot noise with a spread of 0.8 nm. Find the spread of the controlled output. *Checkpoint:* 0.855 nm.

(b) The only disturbance is instead a drift of 0.04 nm per lot. Find the settled offset. *Checkpoint:* 0.160 nm.

1.2 Concept check · 2 min

M-1's after-etch reading of a lot arrives two lots after it, $m = 2$. A step disturbance appears on lot 20. Why are lots 20, 21 and 22 all produced at the full error, whatever the weight λ?

(a) the EWMA averages three readings before it acts on any of them

(b) no reading of lot 20 or later has reached the controller before lot 22 is etched

(c) the trim rate is unknown until three lots have been measured

(d) the integral action of the controller needs three lots to wind up

1.3 Theory and calculation · 5 min

A run-to-run loop adds a fraction of the newest reading's error to its estimate, as in Equation (1.6).

(a) Find the largest stable loop gain K for a queue of one lot and for a queue of three lots. *Checkpoint:* 1.000 and 0.445.

(b) With three lots of queue, the controller runs at $\lambda = 0.4$ but believes a process gain 20 % lower than the true one. Is the loop stable? *Checkpoint:* $K = 0.50$, unstable.

1.4 Theory and calculation · 4 min

M-1's controller uses the matched-lot estimator of Section 1.4.4 with the gain known, $\lambda = 0.5$ and a queue of $m = 3$ lots. A step of 1 nm appears on lot 10. Give the error on lots 13, 14 and 15.

1.5 Concept check · 2 min

Two run-to-run controllers on the same line have the same weight, the same believed gain and the same two-lot metrology queue. One adds a fraction of the newest reading's error to its estimate, and the other regresses each reading on the recipe that made that lot. What separates their stability?

(a) the matched-lot controller samples more lots, so its queue is shorter

(b) the newest-reading controller uses a smaller weight, so it is slower

(c) the matched-lot controller filters the metrology noise, so it is less excited

(d) the matched-lot controller keeps the record of which recipe made each reading, so the delayed term cancels

1.6 Theory and calculation · 5 min

A small plate has three zones in a row with insulated ends. Each zone loses $U = 1\,\mathrm{W/K}$ to the surroundings and exchanges $k_x = 2\,\mathrm{W/K}$ with each neighbor.

(a) Find the steady gain of the uniform pattern, of the pattern with one half-wave and of the alternating pattern. *Checkpoint:* 1.000, 0.333 and 0.143 K/W.

(b) A loop on the center zone is tuned with its neighbors held at their setpoints, to a loop gain of one. What loop gain does it have when all three zones move together? *Checkpoint:* 5.

1.7 Theory and calculation · 3 min

A total-radiation pyrometer is set to an emissivity, but the wafer's true emissivity is 5 % lower. The loop holds the reading at 600 °C. Find the true wafer temperature.

1.8 Theory and calculation · 4 min

A film of 200 nm is etched with a local rate that varies by $\pm 5\,\%$ across the wafer, over an underlayer etched 25 times more slowly. The main etch ends on an endpoint detector that fires when half the wafer has cleared, followed by an over-etch proportional to the main etch.

(a) Find the smallest over-etch that clears every point. *Checkpoint:* 5.26 %.

(b) Find the underlayer loss at the fastest point with 8 % over-etch. *Checkpoint:* 1.02 nm.

1.9 Theory and calculation · 5 min

Lithography leaves a lot-to-lot spread of 1.5 nm in the developed linewidth, which a microscope reads with a spread of 0.5 nm. The etch adds no spread of its own. The etcher's trim is corrected on the same lot from the developed reading, as in Section 1.4.7.

(a) Find the finished spread with full feedforward, $g = 1$. *Checkpoint:* 0.500 nm.

(b) Find the best feedforward gain and the spread it gives. *Checkpoint:* $g_* = 0.900$, 0.474 nm.

(c) With $g = 1$, only one lot in two is measured. Find the spread over all lots. *Checkpoint:* 1.118 nm.

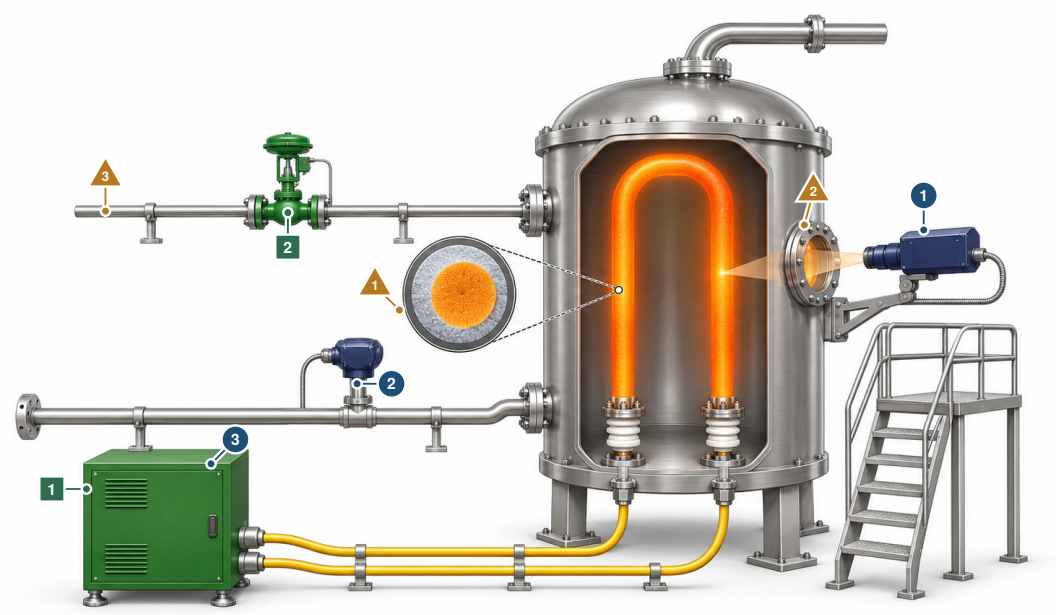
1.10 Concept check · 2 min

A cold wafer lands on a hot bake plate at a known instant, and the plate dips before any thermocouple loop can respond. Which control element answers this disturbance best?

(a) a power pulse triggered by the landing, with the zone integrators held while it runs
(b) a faster integral time on every zone loop
(c) a run-to-run correction of the zone setpoints from the next wafer's linewidth
(d) a filter on the thermocouples to keep the loops from reacting to the dip

CHAPTER 2

Siemens Polysilicon Reactor Simulator for Gain Scheduling and an Unmeasured Constraint



The controller moves 1 rod power supply, current regulated, set by the temperature loop
2 trichlorosilane feed valve, set by the composition loop

The controller reads 1 pyrometer on the rod surface, through a sight glass 2 trichlorosilane in the gas
3 the supply's voltage and current, from which the hidden center temperature is computed

What disturbs it 1 the rods growing thicker, which shrinks the loop gain a hundredfold
2 a silicon film on the sight glass, which makes the pyrometer read low 3 boron in the feed, which no instrument sees

Notice the inset: the rod is hottest at its center, which no instrument reaches, while the pyrometer sees only the surface. Illustration, not to scale.

IN PLAIN TERMS A Heater That Is Also the Product

The main idea. Pure silicon for chips and solar cells grows on thin silicon rods in a sealed jar. A current keeps the rods glowing hot, and a gas leaves new silicon on them, so they grow many times thicker.

The challenge. As the rods thicken, a change in current moves their temperature about a hundred times less. The automatic control set up at the start slowly stops working. Worse, the hottest place is the middle of each rod, where no instrument can look. It can melt while the surface reading looks perfect.

Why it matters. The hidden middle, not the measured surface, decides how large the rods can grow.

THREE WAYS TO READ THIS CHAPTER

Explore	11 min	The first page, the history and fact boxes in Section 2.2, the try-it and fact boxes in Section 2.6, and the key points
Study	60 min	The whole chapter with its pause points and the twelve exercises
Practice	20 min	The first page, Section 2.6, Section 2.7, and Section 2.8 with its open challenge

The control question The silicon rods of a Siemens reactor are the heating element and the product, and they grow from 8 to 150 mm in one batch. How does a temperature loop stay useful while its plant gain falls a hundredfold, and how is the batch ended at a center temperature that no instrument measures?

Learning objectives

- Compute the resistance, current, power and loss slope of a rod pair at a given diameter from the model equations.
- Explain why current regulation is self-regulating and voltage regulation runs away, from the sign of $\partial P/\partial \bar{T}$ against the loss slope.
- Build a gain schedule from a closed-form plant gain, and explain why a schedule quantity with the wrong magnitude and the right shape works.
- Predict the parabolic center-to-surface rise, and protect the unmeasured center with an override on an inferred estimate.
- Judge whether a predictive controller's gain in batch time comes from prediction or from constraint margin.

Prerequisites Sections 1.6.2 (velocity-form PID), 1.6.4 (scenarios and data) and 1.4.2 (what a reactor measures, and a pyrometer that is wrong). Section 1.6.3 is needed only for the predictive controller.

The app https://apmonitor.com/apps/polysilicon-cvd/

See also Volume 7, Chapter 2 for a heat-generation slope that decides stability. Volume 5, Chapter 11 for a soft sensor of a core temperature nothing reaches, and Volume 5, Chapter 10 for overrides on inferred constraints. Volume 4, Chapter 2 for a protective override on a limit.

2.1 Introduction

A silicon rod standing in a Siemens reactor is three things at once. It is the heating element, because the current that holds it near 1100 °C passes down one leg of a hairpin, across a bridge at the top and up the other. It is the deposition surface, because trichlorosilane decomposes on it and leaves silicon behind. And it is the product, because at the end of the batch the reactor is opened and the rods are broken up and sold. The third fact is the one that decides the control problem. A plant whose equipment is fixed and whose operating point drifts can be retuned once. A plant that is its own product changes while it runs, and here it changes by a factor that is not small.

The numbers are set by geometry alone. Resistance is resistivity times length divided by cross-sectional area, and the cross-section grows as the square of the diameter, so growing an 8 mm seed to 150 mm multiplies the area by 352. Measured on the simulator at a surface held at 1100 °C, the rod pair resistance falls from 50.9 Ω to 0.0649 Ω, the current rises from 21.8 A to 2646 A, the terminal voltage falls from 1111 V to 172 V, and the power rises from 24.2 kW to 454 kW. The steady-state gain from current to surface temperature falls from 10.46 °C/A to 0.0828 °C/A, a factor of 126. A proportional-integral controller commissioned in the first hour of the batch is still connected in the last one, and by then it is very nearly open loop.

The shape of the problem is general. Any plant that accumulates or consumes the material it is built around presents the same picture, including a batch crystallizer whose surface area grows, an exchanger whose fouling resistance grows, a fixed bed whose catalyst deactivates, a furnace tube that cokes and a cell whose electrodes age. Gain scheduling is the standard answer and it has a substantial analysis literature, covering when a family of linearizations may be interpolated, what the resulting controller guarantees, and the condition that the scheduling variable must move slowly compared with the closed-loop dynamics [1–3]. What a survey cannot supply is a plant where the schedule is not fitted. Most industrial schedules use a proxy variable, a lookup table and a long commissioning period. Here the scheduling variable is the diameter, it is the cause of the gain change rather than a correlate of it, and the gain itself is a closed-form function of it.

Two further properties come out of the same rod. Silicon is a semiconductor, so its resistivity falls steeply with temperature, and the electrical power therefore falls with temperature at constant current

and rises with temperature at constant voltage. Whether the operating point is stable is decided by which terminal of the same power supply is held constant, and no amount of tuning reaches that question. And heat is generated throughout the rod while it leaves only from the surface, so the radial profile is parabolic and the hottest point is the axis. The pyrometer reads the surface. Silicon melts at the center. The constraint that ends the batch has no instrument on it and never has had one, which is why the rod interior appears in the reactor-modeling literature as a computed field rather than as a measurement [4, 5].

However, a student handed a first-order transfer function and a tuning rule never meets a plant that will not hold still, and a block diagram that draws the power supply as a source of power hides the sign that decides stability. This chapter works with an interactive Siemens reactor in which the loop gain moves by two orders of magnitude during one batch, the regulation mode of the supply decides whether an operating point exists at all, and the binding constraint is available only through a model. Section 2.2 sets out the process and its few instruments. Section 2.3 gives the nine-state deposition and energy balance with its instrument chain and supply model, in the detail needed to rebuild it, and states where each assumption fails. Section 2.4 derives the three properties a controller has to be designed around: the sign that decides stability, a gain that is a closed-form function of the diameter, and a constraint reached only through a model. Section 2.5 describes the scheduled PI loop, the pyrometer, the override on the inferred center and a shrinking-horizon predictive controller of the surface setpoint. Section 2.6 runs the scenario sequence, from a seed rod held by hand to a graded challenge carrying four simultaneous faults. The predictive controller is a negative result there. Under the shipped override it lengthens the batch, and the hour it saves when the override band is narrowed is margin that the override alone recovers. Section 2.7 names the instruments, the supply and the control system behind the loops. Sections 2.8 to 2.10 state the limitations, discuss the energy trade and conclude.

2.2 Polysilicon Production and Its Instrumentation

Silicon for semiconductors and for most photovoltaic wafers begins as metallurgical-grade material at roughly 98 % purity, which is several orders of magnitude short of what a device needs. The industrial route to the remaining purity is chemical rather than metallurgical. The silicon is converted to a chlorosilane, usually trichlorosilane, the chlorosilane is distilled until its impurities are measured in parts per billion, and the purified gas is decomposed back to solid silicon on heated silicon rods [6, 7]. Both stages purify. The distillation does most of the work, and the growth surface rejects impurities again by segregation as the deposit forms. Neither stage is closed-loop within a batch, which is why the industry runs on incoming-material specification and statistical process control for purity rather than on feedback.

> ✦ DID YOU KNOW? **Most of the World's Polysilicon**
>
> Rod deposition from trichlorosilane, in its modern recycling form, makes about 80 percent of the world's polysilicon, and its deposition step typically consumes around 90 kW h for each kilogram, more than half of the energy used in polysilicon production [8]. A complete batch of this chapter's simulator spends 87.0 kW h/kg.

The deposition stage carries the name of the company that commercialized it and the arrangement has proved remarkably durable. Reviews of feedstock technology published decades after its introduction still identify rod deposition from trichlorosilane as the dominant industrial process, with the fluidized-bed silane route as the principal alternative [6, 9, 10]. What has changed is not the geometry but the energy accounting and the modeling. The reactor is among the largest single electricity consumers in the wafer supply chain, and polysilicon production is a substantial term in photovoltaic life-cycle energy analyses [11]. That pressure produced a body of work on where the energy goes, on the radiative exchange between rods and a cooled bell jar wall [12], on the deposition

radius at which the energy per kilogram stops improving [13], and on comparative thermal analysis against the fluidized-bed alternative [5].

> ⧖ HISTORY **Rods of the Same Silicon (1956)**
>
> A patent filed in Germany in June 1956, and granted in the United States in 1961, gave the deposition reactor the form this chapter models [14]. Silicon had been grown from a chlorosilane and hydrogen on a heated strip of tantalum, which then had to be removed from the crust at the risk of new impurities. The patent grew it instead on self-supporting rods of silicon, joined at their upper ends so that the current passed up one rod and down the other. Its rods were 0.5 m long and 3 mm across, held near 1150 °C. The temperature was judged by the glowing color of the rods, and the current was raised by hand with a rheostat, continuously and gradually, because the resistance of the rods fell as they thickened [14]. That falling resistance is the gain change this chapter schedules.

The instrumentation history is short because there is very little of it. The surface temperature is read optically through a sight glass, and everything else the operator can see is electrical. The rod interior has never had an instrument, and it is treated in the literature as a computed quantity. The analysis of temperature homogeneity of polysilicon rods is carried out numerically for exactly that reason [4], and the same is true of the deposition chemistry, where the split between surface kinetics and gas-phase transport of trichlorosilane is established from transport models and growth-rate measurements rather than from anything measured inside the reactor during a run [15]. A reactor with one critical measurement and one critical unmeasured constraint is the natural home of inferential control, which was formalized as the use of secondary measurements to estimate a primary variable that cannot be measured directly [16] and which is now a standard industrial tool under the name soft sensing [17]. The structure that enforces the resulting limit is a selector, and the systematic design of constraint switching with selectors is itself an active subject [18].

2.3 Reactor Model

Reactor R-401 is a single rod pair inside a water-cooled bell jar. Two seed rods of diameter d stand on electrodes in the baseplate and are joined by a bridge at the top, giving an electrical path of length L through solid silicon. Power supply PS-401 drives that path and regulates either its current or its terminal voltage. Trichlorosilane and hydrogen enter at the baseplate, and the exhaust carries away the silicon tetrachloride byproduct. Pyrometer TE-401 reads the rod surface through a sight glass and supplies the only temperature measurement in the machine. Figure 2.1 shows the arrangement, the instrument tags and every symbol used in the equations below.

Table 2.1 classifies what the loops interact with. The clock is in hours, the rod is in millimeters, temperatures are in degrees Celsius, the electrical quantities are in amperes, volts and kilowatts, and the deposit is in kilograms with its boron content in parts per billion atomic. Nine states are integrated. The run stops on its own when the diameter reaches the harvest value or when the computed center temperature reaches the melting point of silicon.

2.3.1 Resistance of a Semiconducting Rod

The resistivity of intrinsic silicon above 1000 °C is dominated by thermally generated carriers, so it falls as the rod heats, and it is represented by a saturating exponential with a high-temperature floor, as shown in Equations (2.1) and (2.2).

$$\rho(T) = \rho_{\infty} + A\,\mathrm{e}^{B/T} \tag{2.1}$$

$$R(\bar{T}, d) = \frac{\rho(\bar{T})\,L}{\pi d^2/4} \tag{2.2}$$

where ρ is in $\Omega\,\mathrm{m}$, T is absolute temperature in kelvin, B is an activation temperature in kelvin, L is the electrical path length of the rod pair in meters and d is the rod diameter. Both sides of (2.2)

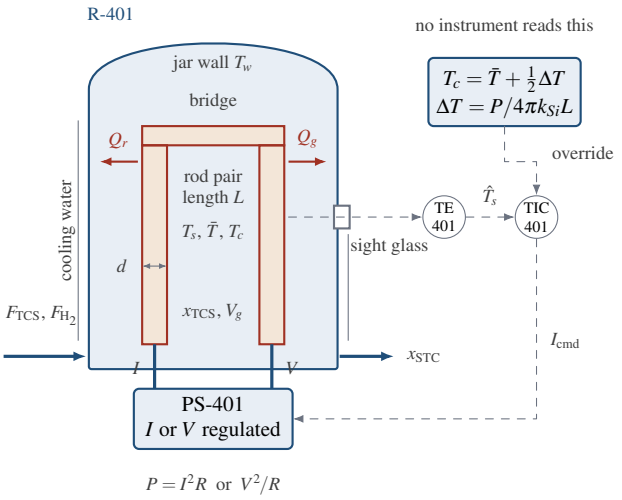


Figure 2.1 Siemens reactor R-401 with one rod pair inside a water-cooled bell jar. Supply PS-401 drives current I at terminal voltage V through an electrical path of length L and diameter d, dissipating P throughout the rod volume. Heat leaves by radiation Q_r to the jar wall at T_w and by convection Q_g to the hydrogen carrier. Pyrometer TE-401 reads the surface $\hat{T}_s$ through a sight glass and TIC-401 manipulates the current command I_{cmd}. The mean rod temperature $\bar{T}$ is the state that is integrated, and the surface T_s and center T_c sit half the parabolic rise ΔT either side of it. The center temperature is computed and never measured.

Table 2.1 Classification of Variables. Nominal Values Are Those of a Seed Rod at the Design Surface Temperature.

Symbol	Role	Nominal	Range
T_s (TE-401)	CV, measured	1100 °C	950–1200 °C SP
T_c	CV, inferred	1117 °C	limit 1150–1414 °C
x_{TCS}	CV, measured	0.120	0.04–0.30 SP
I (PS-401)	MV	21.8 A	0–3000 A
F_{TCS}	MV	2.50 kmol/h	0.2–12 kmol/h
supply mode	DV, operator	current	current or voltage
f_{loss}	DV, unmeasured	1×	0.8–1.3×
v_g gas velocity	DV, unmeasured	1×	0.5–2×
C_B feed boron	DV, unmeasured	0.60 ppba	0.1–4 ppba
f_v vaporizer	DV, unmeasured	1×	0.85–1.15×
σ_T pyrometer noise	DV, measured	2 °C	0–20 °C
b_T pyrometer offset	DV, unmeasured	0 °C	−40–40 °C
d	state	8 mm	8–150 mm
$\bar{T}$	state	1108.4 °C	20–2200 °C
I_a, V_a	states	21.8 A, 1111 V	supply ratings
x_{TCS}, x_{STC}	states	0.120, 0.020	0–0.9
m_{Si}	state	0 kg	0–214 kg
B_{acc}	state	0	≥ 0
E	state	0 kW h	0–18 700 kW h

carry units of ohms. The resistance is evaluated at the mean rod temperature $\bar{T}$ and not at the surface, because the whole cross-section conducts. Two separate effects shrink it over a batch. At a fixed 1100 °C the area term alone takes R from 53.65 Ω to 0.1526 Ω, a factor of 352 that is exactly the area ratio. Holding the *surface* at 1100 °C instead raises $\bar{T}$ from 1108.4 to 1258.0 °C as the rod

thickens, and (2.1) roughly halves again over that span, so the measured resistance falls from 50.9 Ω to 0.0649 Ω, a factor of 785.

2.3.2 Electrical Power and the Energy Balance

The supply regulates one terminal quantity and the rod decides the other, as shown in Equation (2.3).

$$P = \begin{cases} I_a^2 R(\bar{T}, d) & \text{current regulation} \\ V_a^2 / R(\bar{T}, d) & \text{voltage regulation} \end{cases} \tag{2.3}$$

where I_a and V_a are the delivered current and voltage, each of which follows its command through a first-order supply lag τ_{ps} as shown in Equation (2.4).

$$\tau_{ps} \frac{\mathrm{d}I_a}{\mathrm{d}t} = I_{\mathrm{cmd}} - I_a, \qquad \tau_{ps} \frac{\mathrm{d}V_a}{\mathrm{d}t} = V_{\mathrm{cmd}} - V_a \tag{2.4}$$

Under voltage regulation the power is also limited to $I_{\max}^2 R$, which is the supply leaving voltage regulation for its current rating, and the dissipation in either mode is limited to $1.5 P_{\max}$ before it is used, which prevents a numerical excursion from injecting unbounded power during a runaway. The rod is one lumped thermal capacity, and its energy balance is Equation (2.5).

$$\rho_{Si} \frac{\pi d^2 L}{4} c_{p,Si} \frac{\mathrm{d}\bar{T}}{\mathrm{d}t} = 3600 \left(P - Q_{\mathrm{loss}} - q_{\mathrm{dep}} \right) \tag{2.5}$$

where ρ_{Si} is the density of silicon, $c_{p,Si}$ its heat capacity, the product of the three leading factors is the rod heat capacity in J/K, and the factor 3600 converts a watt to a joule per hour so that the derivative is per hour. The rod heat capacity runs from 487 J/K at the seed to 171 kJ/K at harvest. Losses leave the surface by radiation to the jar wall and by convection to the hydrogen carrier, as shown in Equation (2.6).

$$Q_{\mathrm{loss}} = f_{\mathrm{loss}} A \left[\varepsilon \sigma \left(T_s^4 - T_w^4 \right) + h_g \left(T_s - T_g \right) \right], \quad A = \pi d L \tag{2.6}$$

where A is the lateral area of the rod pair in m^2, ε is emissivity, σ is the Stefan-Boltzmann constant, h_g is the convection coefficient to hydrogen, and f_{loss} is the disturbance multiplier on total loss. Temperatures in the radiative term are absolute. Differentiating (2.6) gives the loss slope used throughout the analysis, as shown in Equation (2.7).

$$S \equiv \frac{\partial Q_{\mathrm{loss}}}{\partial T_s} = f_{\mathrm{loss}} A \left(4 \varepsilon \sigma T_s^3 + h_g \right) \tag{2.7}$$

The slope S is 61.6 W/K at the seed rod and 1154 W/K at harvest, growing linearly with d because the area does. The rod thermal time constant $\rho_{Si} \pi d^2 L c_{p,Si} / (4S)$ therefore also grows linearly, from 7.9 s to 148 s.

2.3.3 The Parabolic Radial Profile

Generation is volumetric and loss is superficial, so the rod cannot be isothermal across its section. Uniform generation q''' in a long cylinder of constant conductivity gives a parabolic profile, and the difference between the axis and the surface is Equation (2.8).

$$\Delta T = T_c - T_s = \frac{q''' d^2}{16 k_{Si}}, \qquad q''' = \frac{P}{\pi d^2 L / 4} \tag{2.8}$$

where k_{Si} is the thermal conductivity of hot silicon. Substituting q''' collapses (2.8) to $\Delta T = P / (4 \pi k_{Si} L)$, so the rise depends on the total dissipation and the rod length alone and not explicitly on the diameter at all. That is worth stating carefully, because the d^2 in (2.8) invites the wrong

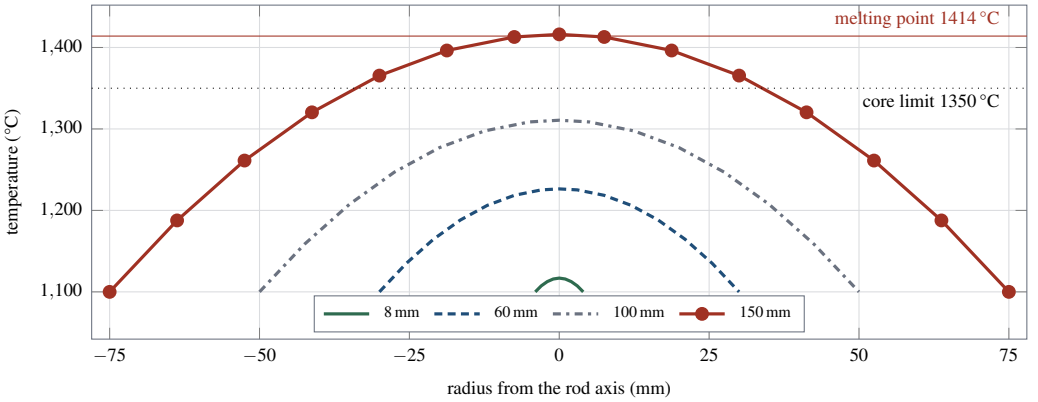


Figure 2.2 Radial temperature profile of the rod at four diameters with the surface held at 1100 °C: 8 mm solid, 60 mm dashed, 100 mm dash-dotted and 150 mm solid with dots. The thin solid line is the melting point and the thin dotted line the core limit. The pyrometer reads the two endpoints of every curve and they are identical. The axis temperature rises from 1116.9 to 1416.1 °C across the same four cases, passes the default 1350 °C limit at 118.7 mm and reaches the melting point at 149.0 mm.

conclusion. At a held surface temperature the power is a surface loss and grows linearly with d, so q''' falls as $1/d$ and the rise grows *linearly*. Measured on the simulator, ΔT is 16.9 °C at 8 mm and 316.1 °C at 150 mm, a ratio of 18.7 against a diameter ratio of 18.75. The area-weighted mean of a parabola over a circular section is exactly halfway between axis and surface, which gives the two algebraic outputs in Equation (2.9).

$$T_s = \bar{T} - \tfrac{1}{2}\Delta T, \qquad T_c = \bar{T} + \tfrac{1}{2}\Delta T \tag{2.9}$$

Only $\bar{T}$ is a state. The pyrometer sees T_s and nothing sees T_c. Figure 2.2 draws the profile at four diameters against the melting point.

2.3.4 Deposition, Gas Phase and Purity

Deposition is a surface reaction in series with transport of trichlorosilane through the boundary layer, and the two resistances add, as shown in Equation (2.10).

$$\frac{1}{k_{ov}} = \frac{1}{k_s} + \frac{1}{k_m}, \quad k_s = k_{s0}\,\mathrm{e}^{-\frac{E_a}{R}\left(\frac{1}{T_s}-\frac{1}{T_{\mathrm{ref}}}\right)}, \quad k_m = k_{m0}\,v_g \tag{2.10}$$

where k_s and k_m are the surface-reaction and mass-transfer growth velocities in mm/h on the radius, E_a/R is an activation temperature, T_{ref} is a reference surface temperature, and v_g is the gas-velocity disturbance multiplier. The diameter and the deposited mass follow in Equation (2.11).

$$\frac{\mathrm{d}d}{\mathrm{d}t} = 2\,k_{ov}\,\frac{x_{\mathrm{TCS}}}{x_{\mathrm{ref}}}, \qquad \dot{m} = \rho_{Si} A\,\frac{1}{2}\frac{\mathrm{d}d}{\mathrm{d}t} \times 10^{-3} \tag{2.11}$$

where the factor 10^{-3} converts millimeters to meters so that $\dot{m}$ is in kg/h. At the design surface temperature and composition the diameter grows at 1.800 mm/h throughout the batch, because neither term in (2.10) depends on d. The reaction absorbs ΔH_{dep} per kilogram, entering (2.5) as $q_{\mathrm{dep}} = \Delta H_{\mathrm{dep}}\dot{m}/3.6$ in watts.

The bell jar gas is one well-mixed volume holding $n_T = V_g/(10^3 v_m)$ kilomoles. Trichlorosilane is fed into a fixed hydrogen carrier, the rods take silicon out of it, and the overall stoichiometry sends four

trichlorosilane molecules to one silicon atom and three tetrachlorides, as shown in Equation (2.12).

$$n_T \frac{\mathrm{d}x_{\mathrm{TCS}}}{\mathrm{d}t} = f_v F_{\mathrm{TCS}} - F_{\mathrm{out}} x_{\mathrm{TCS}} - 4\,\frac{\dot{m}}{M_{Si}} \tag{2.12}$$

$$n_T \frac{\mathrm{d}x_{\mathrm{STC}}}{\mathrm{d}t} = 3\,\frac{\dot{m}}{M_{Si}} - F_{\mathrm{out}} x_{\mathrm{STC}} \tag{2.13}$$

where $F_{\mathrm{out}} = f_v F_{\mathrm{TCS}} + F_{\mathrm{H_2}}$ is the sweep leaving the jar, f_v is the vaporizer calibration disturbance and $M_{Si} = 28.09\,\mathrm{kg/kmol}$. The gas holdup is 0.1786 kmol against a sweep of about 20.5 kmol/h, so the composition time constant is 31.4 s, four orders of magnitude shorter than the batch. Boron arrives dissolved in the feed and is incorporated less selectively as the surface gets hotter, as shown in Equation (2.14).

$$[B] = C_B\, k_{B0}\, \mathrm{e}^{k_{B1}(T_s - T_{\mathrm{ref}})/100}, \qquad \frac{\mathrm{d}B_{\mathrm{acc}}}{\mathrm{d}t} = \dot{m}\,[B] \tag{2.14}$$

where $[B]$ is the instantaneous boron content of the deposit in parts per billion atomic, C_B is the boron in the feed and k_{B1} is expressed per 100 °C. The reported purity is the running average B_{acc}/m_{Si} over the whole deposit, which is what a laboratory would measure on a broken piece of rod. Consumed energy accumulates as $\frac{\mathrm{d}E}{\mathrm{d}t} = P/1000$ in kW h/h.

2.3.5 Parameters, Initialization and Numerics

The constants are collected in Table 2.2. Twenty-two are adjustable in the application and ten are hidden, and the hidden ones are the material properties and reference conditions a reader would have to look up rather than read off the interface.

Every scenario opens at a genuine steady state rather than settling into one. The initialization solves the design point by fixed-point iteration, because the loss is a function of the surface, the resistance is a function of the mean, and the mean sits half the parabolic rise above the surface with a rise that depends on the power. Starting from $\bar{T} = T_{\mathrm{SP}}$, the iteration of Equation (2.15) is repeated eight times.

$$R \leftarrow R(\bar{T}, d), \quad I \leftarrow \sqrt{\frac{Q_{\mathrm{loss}} + q_{\mathrm{dep}}}{R}}, \quad \bar{T} \leftarrow T_{\mathrm{SP}} + \frac{I^2 R}{8\pi k_{Si} L} \tag{2.15}$$

The same expression supplies the manual-mode aid that sets the current for whatever diameter the rod has at that moment. Nine scenarios start the rod part way through a batch at 60, 70, 90, 100 or 140 mm so that the interesting part is reached without waiting for it. In automatic mode the composition loop starts at the trichlorosilane feed that replaces what the carrier sweeps out and what the rods deposit, so the gas opens steady as well.

The integrator is classical fourth-order Runge-Kutta on a fixed step of 0.0004 h (1.44 s). The controller executes every 0.0008 h (2.88 s) and the historian logs 20 samples per simulated hour, one every three minutes. The pseudo-random generator is seeded at 20260919, so a noisy scenario replays identically. Section 1.6.1 describes the engine that integrates, seeds and logs the run.

2.3.6 Assumptions

Each simplification below is deliberate and each one fails somewhere specific.

One rod pair stands for a whole reactor. Production machines carry 36 to 72 rods, and the rods near the middle see each other rather than the cooled wall, so they run hotter than the ones at the edge. Radiative exchange in (2.6) is rod to wall only, with no view factors between rods. The uniformity problem that omission removes is a large part of the real control engineering and it is the subject of the rod-homogeneity literature [4, 12].

The rod is one lumped temperature plus an analytic profile. Equation (2.8) is exact only for uniform volumetric generation and constant conductivity. Neither holds. The conductivity of silicon falls with temperature over this range, which steepens the real profile, and current crowds toward the hotter center because the resistivity there is lower, which concentrates the generation inward [19].

Table 2.2 Model Parameters. Hidden Constants Never Appear on the Panel.

Symbol	Meaning	Value	Unit	Visibility
L	electrical path length	5.2	m	adjustable
d_0	seed rod diameter	8	mm	adjustable
d_{end}	harvest diameter	150	mm	adjustable
k_{Si}	conductivity of hot Si	22	W/(m K)	adjustable
B	resistivity activation temp.	12120	K	adjustable
ε	rod emissivity	0.70	—	adjustable
T_w	bell jar wall	100	°C	adjustable
h_g	convection to hydrogen	60	W/(m^2 K)	adjustable
T_g	bulk gas temperature	350	°C	adjustable
k_{m0}	mass-transfer growth	1.20	mm/h	adjustable
k_{s0}	surface-reaction growth	3.60	mm/h	adjustable
E_a/R	deposition activation temp.	12000	K	adjustable
F_{H_2}	hydrogen carrier flow	18	kmol/h	adjustable
V_g	bell jar gas volume	4.0	m^3	adjustable
k_{B0}	boron incorporation	0.35	—	adjustable
k_{B1}	selectivity loss with T_s	0.55	per 100 °C	adjustable
$[B]_{\text{spec}}$	electronic-grade limit	0.30	ppba	adjustable
$I_{\max}$	supply current rating	3000	A	adjustable
$V_{\max}$	supply voltage rating	1800	V	adjustable
$P_{\max}$	supply power rating	550	kW	adjustable
c_{kg}	value of the product	22	$/kg	adjustable
c_{kWh}	electricity	0.06	$/kW h	adjustable
ρ_∞	resistivity floor	2.0×10^{-5}	Ω m	hidden
A	resistivity prefactor	7.32×10^{-8}	Ω m	hidden
ρ_{Si}	density of silicon	2330	kg/m^3	hidden
$c_{p,Si}$	heat capacity of silicon	800	J/(kg K)	hidden
T_{melt}	melting point	1414	°C	hidden
ΔH_{dep}	heat of deposition	120	kJ/kg	hidden
v_m	molar volume	0.0224	m^3/mol	hidden
τ_{ps}	supply response	0.002	h	hidden
x_{ref}	reference composition	0.12	—	hidden
T_{ref}	reference surface temp.	1100	°C	hidden

Both effects steepen the profile and neither is modeled, so the computed T_c should be read as an estimator that runs low rather than as a truth value.

The gas phase is a single well-mixed volume. There is no axial concentration profile up the rod and no reactor-scale flow field, so the deposition non-uniformity along the rod length that limits real rod quality is absent. The mass-transfer coefficient is a single number scaled by a velocity multiplier rather than a boundary-layer calculation.

The chemistry is one overall stoichiometry. Dichlorosilane, homogeneous nucleation to powder, and the hydrochlorination recycle that converts the tetrachloride byproduct back into feed are all absent, and the last of those is most of a real plant [6]. The byproduct term (2.13) therefore tracks an inventory that has nowhere to go.

Purity is one segregation coefficient. Equation (2.14) lumps species, surface condition and growth rate into two constants, and phosphorus, carbon and metallic contaminants behave differently and are not represented. The model can show that a hotter surface is less selective. It cannot be used to set a specification.

Seed heating and the cold start are absent. Equation (2.1) at room temperature gives a resistance of about 7×10^{15} Ω, so a cold silicon rod is very nearly an insulator and a real reactor has to be struck at kilovolts or preheated before current regulation has anything to hold. Every scenario in the

application begins with the rod already hot.

Disturbances are stylized. A pyrometer offset that appears at once at 25 °C is not how a window fogs, but a step is a fault whose consequence can be read off a trend, while a slow drift would be indistinguishable from the plant. A feed impurity that triples between one sample and the next does not happen in a distillation train either, and it isolates the observation that matters, which is that no instrument in the reactor responds. Scenarios that begin at 60, 70, 90, 100 or 140 mm skip the first two days of a batch, which is defensible only because the diameter is the state that carries all of the history.

The numerics are more repeatable than a control system. A fixed step and a seeded generator reproduce one trajectory on every machine, while a distributed control system delivers jitter, quantization and a scan order that moves.

2.4 Three Properties of the Plant

2.4.1 One Term Decides Stability and Dominates the Gain

Differentiating (2.1) and (2.2) at fixed diameter gives $\partial R/\partial \bar{T} = -RB/\bar{T}^2$ in the exponential-dominated regime, and substituting into (2.3) gives the two slopes of Equation (2.16).

$$\left.\frac{\partial P}{\partial \bar{T}}\right|_I = -\frac{PB}{\bar{T}^2} < 0, \qquad \left.\frac{\partial P}{\partial \bar{T}}\right|_V = +\frac{PB}{\bar{T}^2} > 0 \tag{2.16}$$

A heated body holds an operating point when its dissipation rises with temperature more slowly than its losses do, so the criterion is $\partial P/\partial \bar{T} < S$ with S from (2.7). The same comparison of a generation slope with a removal slope decides ignition in the exothermic reactor of Volume 7, Chapter 2. Under current regulation the left side is negative, the criterion is satisfied by a wide margin, and the rod is self-regulating before any controller is attached. Under voltage regulation the left side is positive and larger than S at every diameter. Evaluated on the simulator at a surface of 1100 °C, the quantity $PB/\bar{T}^2$ is 154, 1068 and 2349 W/K at 8, 60 and 150 mm against loss slopes of 62, 462 and 1154 W/K, so the ratio is 2.50, 2.31 and 2.03. Scenario 5 runs that criterion out on a 60 mm rod sitting at its design point. With the terminal voltage held near 346 V the center reaches the 1414 °C melting point 6.50 min later. With the current regulated instead, starting from the same 525 A, the same rod holds 1100.0 °C for the two hours the case is run. This is a property of the plant and of the material, it is decided entirely by which terminal of one power supply is held constant, and it is invisible in a block diagram that draws the supply as a source of power.

🕮 **Pause.** Try Exercise 2.1 before reading on.

The same term sets the loop gain. Differentiating the steady-state balance $P = Q_{\text{loss}}$ with (2.16) and (2.9) substituted gives Equation (2.17).

$$\frac{\partial T_s}{\partial I} = \frac{2IR}{S + \dfrac{PB}{\bar{T}^2}\left(1 + \dfrac{S}{8\pi k_{Si} L}\right)} \tag{2.17}$$

where the bracket accounts for the fact that raising the power raises the mean temperature by more than it raises the surface. The self-regulation term $PB/\bar{T}^2$ is 2.50 times the loss slope at the seed rod and 2.03 times it at harvest, so it is the larger part of the denominator throughout. Dropping it gives the familiar $2IR/S$, which is the quantity the schedule is built from and which overstates the true gain by a factor of 3.449 at the seed and 3.593 at harvest. The application plots and exports both, labeled `sched_2IR_S` and `gain_dTdI` in the comma-separated record and drawn as two traces on the supply strip. The second is a central difference of the design-point calculation, and it agrees with an independent solution of the model's own equilibrium under a 0.1 % current perturbation to four decimal places at every diameter, giving 10.4645 °C/A at the seed, 0.4240 at 60 mm and 0.0828 at harvest. Equation (2.17) reproduces those to within 7 % over the whole range. Table 2.3 lists the electrical march of the batch and both versions of the gain.

Table 2.3 The Plant at Five Diameters With the Surface Held at 1100 °C and $x_{\mathrm{TCS}} = 0.120$. Every Column Is a Steady State of the Simulator.

d (mm)	8	30	60	100	150
R (Ω)	50.9	3.155	0.659	0.189	0.0649
I (A)	21.8	169.7	525.3	1265	2646
V (V)	1111	535	346	239	172
P (kW)	24.2	90.9	181.7	302.9	454.4
$\bar{T}$ (°C)	1108.4	1131.6	1163.2	1205.4	1258.1
T_c (°C)	1116.9	1163.2	1226.4	1310.7	1416.1
S (W/K)	61.6	230.9	461.7	769.6	1154.3
$PB/\bar{T}^2$ (W/K)	154	558	1068	1679	2349
$2IR/S$ (°C/A)	36.0959	4.6382	1.4986	0.6222	0.2975
$\partial T_s/\partial I$ (°C/A)	10.4645	1.3290	0.4240	0.1741	0.0828
rod τ (s)	7.9	29.7	59.4	98.9	148.4

2.4.2 A Schedule That Is Not Fitted

Every term on the right of (2.17) is a closed-form function of the diameter at a given surface temperature, and the diameter is the integral of a deposition rate the tool already computes. The application therefore schedules the proportional gain by the ratio of the plant gain at the seed to the plant gain now, as shown in Equation (2.18).

$$\gamma(d) = \mathrm{clamp}\left(\frac{g(d_0)}{g(d)}, 0.05, 400\right), \quad g(\delta) = \frac{2I(\delta)R(\bar{T},\delta)}{S(T_s,\delta)} \tag{2.18}$$

where $I(\delta)$ is the design-point current of (2.15) evaluated at diameter δ and the present effective setpoint, R is evaluated at the mean rod temperature and S at the surface temperature that the supply's meters infer (Section 2.5.3), and every term uses the declared parameters of Table 2.2 with the design heat loss and gas velocity, so that only the diameter varies between the numerator and the denominator. The multiplier is referenced to the seed so that the number the reader enters is the gain at the start of a batch. The self-regulation term of (2.17) is deliberately left out of $g(\delta)$, and the reason is the most useful thing this application has to say about gain scheduling. The schedule quantity is wrong in magnitude by a factor of about 3.5 and right in shape, because that factor moves only from 3.449 at the seed to 3.593 at harvest while the schedule quantity itself moves by 121. A schedule needs the shape. The one constant factor it gets wrong is absorbed once, by the proportional gain entered at the seed, and it never has to be revisited. The ratio of the simplified expression between seed and harvest is 121 against 126 for the measured derivative, a difference of 4 % over two orders of magnitude.

🕮 **Pause.** Try Exercise 2.2 before reading on.

Gain scheduling carries a standing warning that the scheduling variable must vary slowly against the closed-loop dynamics [1, 2]. Here that condition is met by a margin that is almost embarrassing. The rod time constant is between 7.9 and 148 s while the diameter takes 79 h to traverse its range, so the scheduling variable moves three orders of magnitude more slowly than the plant it schedules. It is also monotonic, unambiguous, and the physical cause of the gain change rather than a correlate of it [3].

2.4.3 A Constraint Reached Only Through a Model

Combining (2.8) and (2.9) gives a center temperature computed from the electrical power and two geometric constants. With the surface held at 1100 °C the center passes the default 1350 °C limit at 118.7 mm and reaches the 1414 °C melting point at 149.0 mm, one millimeter short of the harvest diameter. Dropping the surface setpoint to 1050 °C moves the 1350 °C crossing out to 162 mm, past harvest. The recipe for the last third of a batch is therefore a constraint problem rather than a tuning

problem, and the variable it is written on cannot be measured. Its estimator is the model in (2.8) applied to the supply's meters, with the mean temperature taken from the measured resistance V/I through the declared resistivity law of (2.1) and the rise taken from the measured power VI at the declared k_{Si} and L, which is the classical inferential arrangement of a primary variable reconstructed from secondary measurements [16, 17]. The kiln of Volume 5, Chapter 11 estimates the core of a brick that no instrument reaches in the same way.

2.4.4 A Reaction on the Flat Part of an Arrhenius Curve

The two resistances in (2.10) are not comparable at the operating point. Evaluated at $x_{\mathrm{TCS}} = 0.120$ and $v_g = 1$, the surface reaction delivers $k_s = 1.812$, 2.588, 3.600 and 5.190 mm/h at surface temperatures of 1000, 1050, 1100 and 1160 °C, against a mass transfer of 1.20 mm/h that does not depend on temperature at all. Above about 1050 °C the transport step is the slower one and it sets the rate. Raising the surface from 1100 to 1160 °C therefore raises the diameter growth from 1.800 to 1.949 mm/h, a gain of 8.3 %, while the boron content of (2.14) rises from 0.2100 to 0.2921 ppba, a loss of 39 % against a specification of 0.30 ppba. The same sixty degrees adds 34 °C to the parabolic rise, so the center at 100 mm moves from 1311 to 1405 °C. Every one of those three consequences is available from the model before the application is opened, and none of them is visible in a rate equation read without asking which of two series terms is smaller.

2.5 Controller Design

2.5.1 The Implemented Algorithm

TIC-401 and the composition controller are both instances of the velocity-form PID of Section 1.6.2, with the derivative taken on the measurement. Equations (2.19) to (2.21) write it with the schedule multiplier of this plant.

$$e_k = a\left(T_{\mathrm{eff},k} - \hat{T}_{s,k}\right) \tag{2.19}$$

$$P_k = K_c\,\gamma_k\,e_k, \qquad D_k = -K_c\,\gamma_k\,\tau_D\,a\,\frac{\mathrm{d}\hat{T}_s}{\mathrm{d}t}$$

$$\Sigma_k = \Sigma_{k-1} + (P_k - P_{k-1}) + (D_k - D_{k-1}) \tag{2.20}$$

$$\begin{aligned} c_k &= u_{\mathrm{bias}} + \Sigma_k + I_{k-1} \\ u_k &= \min(I_{\max},\ \max(0,\ c_k)) \\ I_k &= I_{k-1} + \left[\frac{K_c\,\gamma_k}{\tau_I}e_k + \frac{u_k - c_k}{T_t}\right]\Delta t_c \end{aligned} \tag{2.21}$$

where $a = +1$ for both loops because both process gains are positive and a rod reading below setpoint calls for more current, which is reverse action in the convention of Section 1.6.2, γ_k is the schedule multiplier of (2.18) and equals one when scheduling is off, T_{eff} is the setpoint after any override, and Δt_c is the controller period. The symbols map onto Chapter 1 as $\Sigma_k = S_k$, $c_k = v_k$ and $u_{\mathrm{bias}} = u_b$. The running sum is never clipped, and the back-calculation anti-windup, with $T_t = \tau_I$, acts on the integral alone, as Section 1.6.2 describes. It sees only the limits of (2.21). Under voltage regulation the current command is converted to a terminal voltage $V_{\mathrm{cmd}} = u_k R_{\mathrm{ref}}$, with R_{ref} the resistance at the design point for the present diameter and effective setpoint, and clamped at the 1800 V rating, and the supply caps the delivered power at $I_{\max}^2 R$ and $1.5 P_{\max}$. None of these is known to the controller, so while one binds the integral keeps winding, as Section 2.6.2 shows. Both loops run with $\tau_D = 0$, so the derivative term is present in the implementation and inactive in every scenario, and no rate limit is configured on either output. The temperature output is clamped to 0–3000 A and the feed output to 0.2–12 kmol/h. Transfer between Manual and Auto is bumpless in both directions, as in Section 1.6.2. The controller period is 2.88 s against a physics step of 1.44 s, and both are far shorter than the shortest rod time constant of 7.9 s.

The default tunings are $K_c = 0.0080\,\mathrm{A/°C}$ and $\tau_I = 9.0\,\mathrm{s}$ on temperature, and $K_c = 10$ kmol/h per mole fraction with $\tau_I = 31.3\,\mathrm{s}$ on composition. The composition reset time is set equal to the 31.4 s gas residence time of (2.12).

2.5.2 The Instrument Chain

The pyrometer is the only sensor in the loop. Its reading is formed from the true surface temperature by adding a Gaussian sample scaled to σ_T, adding the fixed offset b_T, and passing the sum through a first-order filter of time constant τ_p, as shown in Equation (2.22).

$$\begin{aligned} \hat{T}_{s,k} &= \alpha\,\hat{T}_{s,k-1} + (1-\alpha)\left(T_{s,k} + \sigma_T n_k + b_T\right) \\ \alpha &= \mathrm{e}^{-\Delta t_c/\tau_p} \end{aligned} \tag{2.22}$$

where n_k is a unit normal deviate drawn once per controller execution and τ_p defaults to 1.8 s. The exact zero-order-hold update is unconditionally stable at any sample period. The offset b_T is the disturbance that matters, because an emissivity error or a silicon film on the sight glass is a bias and not noise, and a bias inside the loop is indistinguishable from a setpoint change. Section 1.4.2 shows how an emissivity error becomes such a bias in a pyrometer reading.

2.5.3 The Override on the Inferred Center

The center-temperature limit is enforced by retreating the surface setpoint inside a band Δ below the limit, as shown in Equation (2.23).

$$T_{\mathrm{eff}} = \begin{cases} T_{\mathrm{SP}} & \mu \geq \Delta \\ T_{\mathrm{SP}} + K_{ov}\,(\mu - \Delta) & \mu < \Delta \end{cases}, \quad \mu = T_{c,\max} - T_c \tag{2.23}$$

where μ is the margin to the limit in degrees, Δ defaults to 80 °C and the override gain K_{ov} defaults to 1.0 °C/°C. The structure is a low selector written as a continuous retreat rather than as a switch, so the two objectives blend over the band instead of trading control of the actuator abruptly [18]. The compressor of Volume 4, Chapter 2 guards its surge limit with an override of another form, a loop whose output rests on its clamp until the limit approaches. The action here is proportional only, so the center settles a little short of the limit rather than exactly on it and the band is what decides by how much. A first-order override needs a band wide enough to arrest an approach that is itself accelerating, and Section 2.6.4 measures what 40, 80 and 120 °C each cost. The important property is the independence of the two paths. The override is computed from the supply's voltage and current and the diameter, the mean temperature from the resistance V/I through the declared resistivity and the rise from the power VI through (2.8) at the declared k_{Si} and L, the pyrometer appears nowhere in that calculation, and a fault in the temperature measurement therefore shows up in the override as a constraint that is moving for no visible reason.

Every quantity the controllers compute is one the tool could know. The override, the schedule of (2.18), the open-loop current, the voltage reference, the starting trichlorosilane feed and the manual-mode aid of (2.15) are evaluated with the declared parameters of Table 2.2, the design heat loss and gas velocity, and the temperatures the supply's meters infer. None of them reads the plant's own mean rod temperature or the run's heat-loss factor, gas velocity and parameters, which no instrument on a real reactor reports. At the declared values the metered center equals the plant's own to within 1×10^{-6} °C. A changed k_{Si}, resistivity activation temperature B or rod length L is therefore a model error that the override cannot see, and Section 2.6.5 measures how far each one moves the estimate.

2.5.4 A Shrinking-Horizon Predictive Controller

A model predictive controller (MPC) is provided as a third structure for the temperature loop, beside the open loop and the PI on the pyrometer. It moves one variable, the setpoint u of the surface loop,

and leaves the PI with its schedule underneath and the override of (2.23) above, so the override retreats the planned setpoint exactly as it retreats the operator's. The problem it addresses is the one of Section 2.4.3 posed as a batch optimization, in which growth is the objective, the surface setpoint trajectory is the decision and the center temperature is a path constraint [20], with growth itself as the objective in place of a tracking error [21]. Studies of the Siemens reactor have tailored the rod temperature profile for throughput in the same spirit [22]. It departs from the linear model and the tracking objective of the option most apps share (Section 1.6.3), as the following paragraphs set out.

The controller's state is the diameter instrument and the center estimate of Equation (2.24), the inferential calculation of Section 2.5.3 written out.

$$\hat{T}_c = \frac{B}{\ln[(\rho_m - \rho_\infty)/A]} + \frac{VI}{8\pi k_{Si} L}, \qquad \rho_m = \frac{V}{I}\,\frac{\pi d^2}{4L} \tag{2.24}$$

where V and I are the supply's voltmeter and ammeter readings, ρ_m is the measured resistivity, the first term inverts (2.1) and (2.2) for the mean rod temperature in kelvin, and every constant takes its declared value. The pyrometer does not enter it. Over the rod of scenario 7 from 100 to 150 mm the estimate departs from the true center by less than 1×10^{-6} °C at the declared values, by −21 to −30 °C with $k_{Si} = 18$, by +16 to +22 °C with $k_{Si} = 26$, by −134 to −139 °C with B raised 10 %, and by −5 to −11 °C with rods 6.0 m long.

The controller's model is not identified from step tests. It is the design point of (2.15) at the declared parameters and the design disturbances, which gives the center $f(d,u)$ and the current $I(d,u)$ at a diameter and a surface temperature, together with the diameter growth $g(u)$ of (2.11) at the composition setpoint. It assumes the surface loop holds its setpoint, which the scheduled PI does within minutes while the diameter moves less than 0.09 mm between decisions. Two output biases make it offset-free, $b_T = \hat{T}_c - f(d,\hat{T}_s)$ and $b_I = I - I(d,\hat{T}_s)$, each filtered by a first-order lag of 0.01 h. They absorb what the model leaves out, such as a heavier heat loss, a slower gas or a fogged sight glass, but they cannot absorb an error in $\hat{T}_c$ itself.

A decision is taken every 60 controller executions (0.048 h, counted in integers), and at once when b_T has moved 3 °C or b_I 2 % of the rating since the last one. The horizon is the time the previous plan predicts the rod needs to reach d_{end}, at most 120 h, cut into at most 16 graduated blocks. The first two last one decision period each, every later block is as long as the two before it together, and the last takes the remainder, so the horizon shrinks with the batch. Measured on scenario 11 it is 13 blocks and 28.55 h at 100.5 mm, seven blocks and 1.20 h at 148 mm and two blocks and 0.10 h at 149.95 mm. Each decision solves Equation (2.25) for the block setpoints u_j and two slacks.

$$\begin{aligned}
\min_{u,s_T,s_I}\ & -\sum_{j=1}^{N_b} \frac{\tau_j\, g'(\bar{u}_j)}{g(T_{\text{SP}})}\,(u_j - \bar{u}_j) + S\sum_{j=1}^{N_b}(u_j - u_{j-1})^2 \\
& \quad + \rho\left(s_T^2 + s_I^2\right) + \sigma\,(s_T + s_I) \\
\text{s.t. } & 950\,^\circ\text{C} \le u_j \le T_{\text{SP}}, \qquad |u_j - \bar{u}_j| \le 40\,^\circ\text{C} \\
& f(d_j,u_j) + b_T \le T_{c,\max} - 5\,^\circ\text{C} + s_T \\
& \frac{I(d_j,u_j) + b_I - 0.98\,I_{\max}}{0.01\,I_{\max}} \le s_I, \qquad s_T,\ s_I \ge 0 \\
& d_j = d_{j-1} + \tau_j\, g(u_j)
\end{aligned} \tag{2.25}$$

where τ_j is the length of block j, $\bar{u}_j$ is the previous plan at that block, u_0 is the setpoint in force, d_0 is the measured diameter, and d_j is capped at d_{end} so that the center is judged only up to the harvest. The objective is in hours. The first term is the diameter the plan adds, expressed as hours of growth at the operator's setpoint T_{SP}, which is also the ceiling because a hotter surface is less pure. The move weight is $S = 10^{-4}$ h/°C^2, and the slacks, in degrees for the center and in percent of the rating for the current, are priced at $\rho = 10$ and $\sigma = 10$ hours per unit squared and per unit. A degree of setpoint is

Table 2.4 Scenario Sequence and the Mechanism Each Teaches. The Starting Diameter Is 8 mm Unless Shown.

#	Scenario	What it teaches
1	Seed rod by hand, MANUAL	arithmetic of a 51 Ω heating element
2	Loop closed on a small rod	closed-loop step on a fast plant
3	Same tuning at 140 mm	a plant gain 111 times smaller
4	Schedule on diameter, 140 mm	closed-form gain scheduling
5	Voltage regulation, 60 mm	thermal runaway to the melting point
6	Run to harvest, 100 mm	the unmeasured center constraint
7	Override on, 100 mm	constraint handling on an estimate
8	Surface at 1160 °C, 60 mm	rate against purity
9	Sight glass fogs, 90 mm	sensor bias inside the loop
10	Feed boron triples, 70 mm	a fault no instrument sees
11	MPC under the override, 100 mm	a plan on an estimated limit
12	Graded challenge, 100 mm	four faults at once

worth 0.048 to 0.057 h of batch time over a 30 h batch, so the linear price is about two hundred times larger and the soft limits hold exactly whenever they can be held. The constraint rows are linearized along $\bar{u}$ by central differences of 0.5 °C and 0.5 mm, the condensed program of $N_b + 2$ variables and at most $4N_b + 2$ rows is solved by the dual active-set method [23], and the rows are relinearized along the new plan up to three times, stopping when no block moves more than 0.05 °C. Scenario 11 took two passes in 634 of its 642 decisions and three in the other eight, the nonlinear model’s center along every accepted plan stayed at or below the 1345 °C ceiling, and the largest slack was zero.

Only the first block is applied. Acceptance requires an optimal status from every pass, finite values of the applied move, the planned trajectory, its predicted center and current, its cost and both slacks, planned setpoints within their bounds, and the expected number of blocks. A rejected plan is marked in the historian and the request returns to the operator’s setpoint over a 0.25 h ramp, and the next accepted plan ramps back over the same interval. Five injected faults, one for each acceptance check, were each rejected with the reason stated and handed over with a largest request step of 0.36 °C per execution. The plan does not model the override above it, and its prediction of the hours to harvest and of the highest center is therefore its own outcome rather than that of the trimmed run. Measured in the browser on a shared and loaded host, the controller with its decisions takes between 6.2 and 6.7 % of one processor core at the fastest simulation speed and under 0.8 % at the default speed, with a mean decision of 0.15 to 0.41 ms and a largest of 1.9 ms.

2.6 Results

Table 2.4 lists the twelve scenarios, which are ordered as a laboratory sequence. A scenario fixes the starting mode, the initial diameter, the two tunings, the control options, the disturbance settings and any timed events.

Scenario 1 makes the point with arithmetic before any controller is involved. The manual aid returns 21.8 A for an 8 mm rod at 1100 °C, which is 1111 V across 50.9 Ω for 24.2 kW. An hour later the rod is 9.66 mm, the same 21.8 A holds only 1037.0 °C, and the calculation has to be done again. At harvest the same rod pair needs 2646 A at 172 V, so a supply for this machine has to deliver a kilovolt at tens of amps on the first day and a quarter of a kilovolt at more than two thousand on the fourth.

2.6.1 The Same Controller on a Plant That Grew

Scenarios 3 and 4 apply an identical 20 °C setpoint step to a 140 mm rod, differing only in whether (2.18) is active. With scheduling off, the surface reaches 1053.2 °C two and a half hours after a step to 1070 °C and never enters a ±2 °C band, leaving 16.8 °C of error outstanding. With scheduling on, the multiplier γ is 50.9 at the controller scan that applies the step, on a rod that has

Table 2.5 Four Complete Batches From an 8 mm Seed With Identical Tuning. IAE Is Accumulated Against the Effective Setpoint Over the Whole Run, Whose Length Is Given in the First Row.

	PI only	PI + sched.	PI + ovr.	PI + both
Batch length (h)	80.14	78.57	81.55	80.90
Final diameter (mm)	150.0	149.3	150.0	150.0
Rod melted	no	**yes**	no	no
Peak T_c (°C)	1387	1414	1317	1323
Final T_s (°C)	1083.8	1099.7	1043.7	1047.1
IAE (°C h)	760.4	43.1	625.6	38.8
Max. deviation (°C)	16.16	2.24	10.31	2.24
Current travel (A)	2432	2596	2001	2035
Silicon (kg)	213.5	211.5	213.5	213.5
Energy (kW h/kg)	87.5	88.4	86.6	87.0
Boron (ppba)	0.1972	0.2095	0.1826	0.1873

reached 140.8 mm, and the true surface settles inside the band in 331 s and overshoots by 0.40 °C, or 2.0 % of the step. The overshoot the scorecard reports from the logged pyrometer reading is 15.6 %, because the 2 °C pyrometer noise is a tenth of the step, and over six noise seeds it ranges from 11.7 to 22.7 % while the settling time ranges only from 330 to 359 s. Read at $t = 4$ h, the integrated absolute error is 65.0 °C h unscheduled against 5.9 °C h scheduled, a factor of 11.0, and the actuator travel rises from 206 A to 420 A. The loop was never unstable and was never mistuned. It became slow because the plant became sluggish underneath it.

> **TRY IT IN THE APP** **3 · The Same Tuning at 140 mm**
>
> Load it and press Run. At half an hour the setpoint steps from 1050 to 1070 °C, and two and a half hours later the unscheduled loop has the surface only at 1053.2 °C. At 4 h the scorecard's IAE reads 65.0 °C h. Load 4 · Scheduling on a Diameter the Tool Knows Exactly and run it to the same time: the step settles within minutes and IAE reads 5.9 °C h.

Running the whole batch separates the two further. Table 2.5 reports four complete runs from an 8 mm seed to harvest or failure, with the same PI tuning in every column and only the two option switches changed.

The first column is a loop that slowly gives up. The diameter is a ramp disturbance on the current requirement, a proportional-integral loop tracks a ramp with a standing offset inversely proportional to its gain, and the gain is falling, so the surface droops monotonically from 1096.2 at 20 mm to 1084.8 °C at 140 mm. The second column never lets the surface fall more than 2.24 °C below setpoint, a deficit it reaches in the first hour on the seed rod, holds it between 1099.0 and 1099.7 °C from 21 mm to the end of the batch, and cuts the accumulated error by a factor of 17.6 over runs of 80.14 and 78.57 h. Figure 2.3 draws the accumulated error and the peak center of the four batches side by side.

The second column also melts the rod. Holding 1100 °C accurately all the way to 149.3 mm is exactly what drives the unmeasured center to 1414 °C, while the drooping surface of the first column keeps the peak center at 1387 °C and survives. Improving the loop made the outcome worse, and the reason is that the regulated variable was not the one that ended the batch. The same situation is visible part way through scenario 6, where a pyrometer reading 1100.2 °C against a 1100 °C setpoint accompanies a computed center of 1393.4 °C, as shown in Figure 2.4. Run on to the end, scenario 6 takes the center past the 1350 °C limit at 118.9 mm and melts the rod at 149.2 mm, within a few tenths of a millimeter of the steady-state crossings of Section 2.4.3. Adding the override of Section 2.5.3 to the scheduled loop caps the peak center at 1323 °C, harvests the full 150 mm in 80.90 h and yields 213.5 kg at 87.0 kW h/kg. The override costs 2.34 h of batch time against the run that melted, which

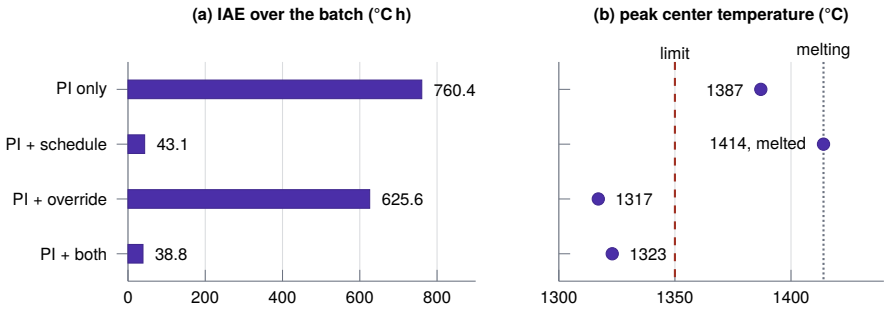


Figure 2.3 The four complete batches of Table 2.5, with identical tuning. (a) The integrated absolute error over the whole batch: the schedule cuts it from 760.4 to 43.1 °C h, and from 625.6 to 38.8 °C h with the override in service. (b) The peak true center temperature against the 1350 °C limit (dashed) and the 1414 °C melting point (dotted). Notice that the run that tracks its setpoint best without the override is the one that melts: the regulated variable is not the one that ends the batch.

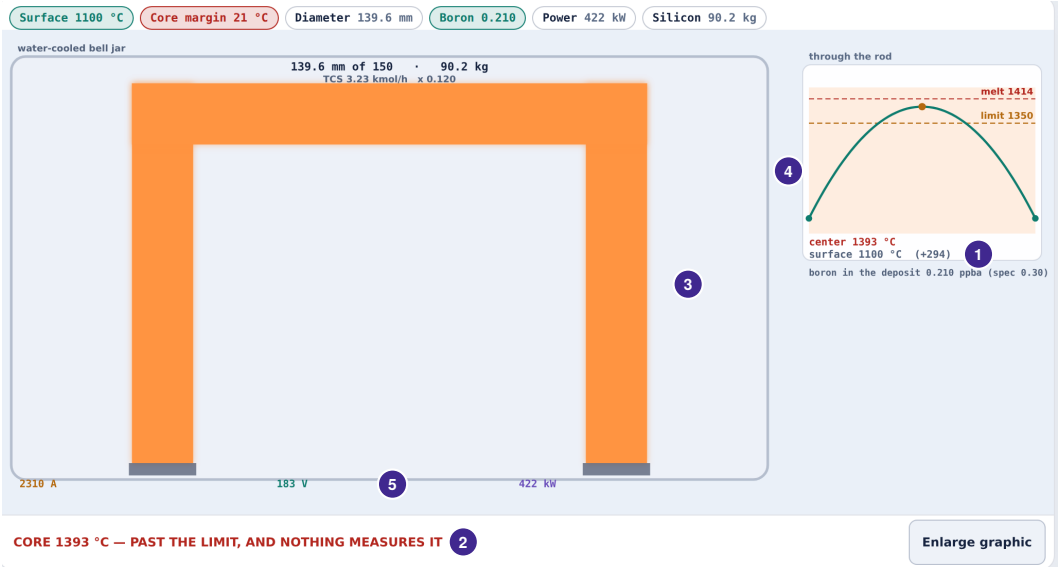


Figure 2.4 Scenario 6 at $t = 22.0$ h, with the center-temperature override out of service, cropped to the plant view. The scheduled loop holds the pyrometer at 1100.2 °C against a setpoint of 1100 while the center stands at 1393.4 °C, past the 1350 °C limit and 21 °C from the melting point, on a rod that has reached 139.6 mm (3). The cut through the rod (4) draws the parabola of Figure 2.2 against the limit and the melting point, with the surface the pyrometer reports at its two ends (1) and a center-above-surface rise of 293.8 °C. The alarm line (2) states what the panel knows: the center is past the limit and nothing measures it. The supply's readings at the foot of the rods (5), 2310.0 A at 182.83 V for 422.35 kW, are what the override computes the center from.

buys the last 0.7 mm, a center that is still 91 °C from the melting point, and a deposit at 0.1873 ppba instead of 0.2095 ppba.

DID YOU KNOW? The Loop That Tracks Worse Uses Less Energy

From the simulator: in Table 2.5 the unscheduled loop with the override accumulates an error of 625.6 °C h, against 38.8 °C h with the schedule added. Yet it harvests the same 213.5 kg with the least energy of the four, 86.6 kW h/kg, and the least boron, 0.1826 ppba. Its surface ends the coolest, at 1043.7 °C.

2.6.2 Losing the Rod to the Choice of Terminal

Scenario 5 starts a 60 mm rod at its design point and runs it open loop, with the same steady-state calculation supplying the command under each regulation mode. Under current regulation the surface holds 1100.0 °C for the whole two hours, and the only thing that moves is the drift the growing rod causes, with the current rising from 525 to 579 A and the terminal voltage falling from 346 to 333 V as the deposit thickens. The power slope logged in the exported record (columns `dpdt_W_K` and `loss_slope_W_K`) is −1068 W/K against a loss slope of 462 W/K.

Under voltage regulation nothing else is changed and the rod is destroyed. The terminal voltage is held between 345 and 346 V, the surface leaves 1100 °C within four minutes, and the center reaches the 1414 °C melting point 6.50 min after the run starts. At that instant the surface is 1172.6 °C, the current has risen on its own from 525 to 1008 A and the power from 182 to 348 kW, all at a voltage that barely moved. That is the positive slope of (2.16) closing on itself, and the slope logged before anything happened was +1068 W/K against the same 462 W/K. The application stops the run at the melt. Continuing the integration past it, as a counterfactual model run, the delivered power reaches the 825 kW ceiling of $1.5P_{\max}$ at 6.91 min and the current reaches the 3000 A supply rating at 7.10 min, 0.6 of a minute after the center passed the melting point. The supply has left voltage regulation, the terminal voltage collapses from 345 to below 225 V by 7.4 min, and the surface peaks at 1678 °C and stands at 1618 °C, with the axis at 2047 °C, two hours in. The saturation flag is raised from the controller scan at 7.01 min onward, which is worth noticing, because the controller never asked for the rating. Its command stays between 528 and 595 A while the rod pulls the rest on its own, so the flag means the actuator is out of room rather than that the loop wanted more.

TRY IT IN THE APP 5 · Current Against Voltage

Load the scenario and press Run. The supply holds the voltage of a 60 mm rod at its design point, and no controller acts. Watch the center trace climb until the run stops itself, the rod melted 6.50 min after the start, while the current has risen on its own from 525 to 1008 A. Reset, set Power supply regulates to Current and run again: the surface holds 1100.0 °C.

Closing the temperature loop does not rescue it, and the way it fails is worth a sentence. The pyrometer sees the surface leave its setpoint and the controller raises the current command, which raises the voltage command, but the plant underneath is unstable in both directions and this scenario departs downward. The surface falls through 1085.1 °C at 3 min, 952.3 °C at 4.8 min and 521.2 °C at 9 min while the resistance of (2.1) climbs through its exponential, so $P = V^2/R$ falls the whole way. The voltage command reaches the 1800 V supply ceiling at 59.8 min, and because that clamp is applied after the controller, the integral keeps winding until the current command itself reaches 3000 A at 65.6 min. At 6 h the rod sits at 294.4 °C drawing 0.0 A. Feedback wrapped around an unstable plant whose actuator authority collapses with the state is not a fix, and no tuning changes the outcome.

2.6.3 A Bias Inside the Loop and a Fault With No Instrument

Scenario 9 deposits silicon on the sight glass at $t = 2$ h so the pyrometer reads 25 °C low. With the override off, the loop does precisely what it was asked to do. Two hours later the reading is 1097.6 °C, within the pyrometer noise of a setpoint it appears to be holding, while the true surface has risen from 1099.7 to 1124.3 °C and the true center from 1297 to 1343 °C. By $t = 20$ h the center stands at 1410 °C, four degrees below melting, and the boron average has risen from 0.2096 to 0.2378 ppba.

With the override on the rod is already being held back before the fault arrives, at 1089.1 °C rather than 1099.7, and two hours after the same event the true surface is only 6.8 °C higher, because the override sees the extra power and retreats the effective setpoint to 1071.3 °C. The center ends the run at 1321 °C instead of 1410 °C, and the deposit at 0.1947 ppba instead of 0.2378 ppba, for 72.1 kg against 76.9 kg. The estimator catches the sensor fault precisely because the faulty sensor is not one of its inputs. An inferred variable built from an independent set of measurements is a cross-check as well as a constraint handler, and on a plant with one critical instrument it is often the only cross-check available.

Scenario 10 is the case with no such luck. The upstream distillation lets three times the boron through at $t = 2$ h. The true surface temperature reads 1099.63 just before the event and 1099.38 °C two hours after it, which is the wander the pyrometer noise causes. The gas composition stays at 0.1200, the growth rate stays at 1.799 mm/h, and the current rises only as the diameter dictates. The boron in the deposit goes from 0.2096 to 0.629 ppba, twice the 0.30 ppba specification, and the average over the whole deposit reaches 0.5677 ppba by $t = 20$ h. No instrument in the reactor responds, the product is scrap, and the only way to establish it is to analyze a piece of the rod after the batch.

🕮 **Pause.** Try Exercise 2.3 before reading on.

2.6.4 Rate Against Purity, and the Graded Challenge

Scenario 8 runs a 60 mm rod with the surface setpoint at 1160 °C for throughput and the override in service. Over 20 h it produces 56.0 kg at 90.8 kW h/kg with a boron average of 0.2414 ppba. The same scenario at 1080 °C produces 51.2 kg at 86.9 kW h/kg with 0.1877 ppba. Eighty degrees of extra setpoint bought 9.4 % of mass, cost 4.5 % in energy per kilogram and 29 % more boron, and moved the center from 1270 to 1320 °C. The surface does not stay at 1160 °C, because the override trims the effective setpoint to 1144 °C within three minutes and takes the surface to 1109.7 °C by 20 h as the center approaches the limit, so the measured trade is smaller than the 8.3 % against 39 % of Section 2.4.4 only because the constraint intervened first.

The challenge starts a 100 mm rod with four things wrong at once. The loop is unscheduled, the setpoint is 1140 °C instead of 1100, the bell jar is losing 10 % more heat than the recipe assumed, and the sight glass fogs at $t = 6$ h. The target is to run the rod to the 150 mm harvest with the measured surface within 25 °C of its effective setpoint and the rod's true center below 1350 °C from the first hour on, and the deposit inside specification at harvest. Changing a model parameter or a disturbance, or firing an extra disturbance, voids an attempt. Run as configured, the rod melts at $t = 6.93$ h and 113.0 mm, with a peak deviation of 42.2 °C between the logged pyrometer reading and its setpoint. The suggested configuration turns scheduling and the override on and takes the setpoint back to 1100 °C. It reaches harvest at 29.7 h with a worst surface error of 5.4 °C after the first hour, an IAE of 39.7 °C h over the run, 118.95 kg deposited, and a boron average of 0.171 ppba against the 0.30 ppba specification. As loaded the center stands at 1396 °C, past the limit, but settings chosen before Run set the starting state, so the suggested run starts from the 1100 °C design point with the center at 1332 °C. The override pulls it down to 1293.5 °C about six minutes in, it stays below 1303 °C through the first six hours, and drifts up as the rod thickens, to 1342 °C at 24 h and a peak of 1349.0 °C just before harvest, 1.0 °C inside the limit. The completion rule judges the center on the logged record from the first hour on, where its peak is 1348.8 °C, 1.2 °C inside the limit, and the worst surface error is 19.6 °C inside its band. Over six pyrometer-noise seeds the judged center peak stays between 1348.6 and 1348.9 °C and every run passes. The width of the override band is what decides that margin, because the action of (2.23) is proportional and the offset grows with the diameter. Narrowing Δ from the default 80 °C to 40 reaches harvest 0.83 h sooner but lets the center peak at 1374.8 °C, 24.8 °C past the limit, and leaves it above the limit for 18.07 h, so the attempt fails. Widening Δ to 120 °C holds the peak to 1323.3 °C and costs 0.94 h of batch time. Because every run ends at the

Table 2.6 The Predictive Controller Against the Override on the Rod of Scenario 7 (100 mm to 150 mm, Same Seed, Noise and Instruments). The Margin Is the Least Distance of the True Center Below the 1350 °C Limit, and the Band Δ and Gain K_{ov} Are Those of (2.23).

	Structure	Batch (h)	kW h/kg	Peak T_c (°C)	Margin (°C)	Boron (ppba)	At I_{max} (h)	Solves
Standard recipe, 3000 A supply								
1	PI and override, $\Delta = 80$ (shipped)	29.595	85.99	1323.0	27.0	0.172	0	—
2	MPC under the override, $\Delta = 80$ (scenario 11)	29.940	85.76	1310.3	39.7	0.167	0	642
2′	MPC with the band narrowed to $\Delta = 5$	28.482	87.28	1345.6	4.4	0.193	0	609
3a	Override retuned, $\Delta = 50$, $K_{ov} = 1$	28.956	86.64	1342.0	8.0	0.183	0	—
3b	Override retuned, $\Delta = 20$, $K_{ov} = 4$	28.675	87.02	1340.9	9.1	0.189	0	—
4a	MPC, $\Delta = 80$, without the center row	29.595	85.99	1323.0	27.0	0.172	0	628
4b	MPC, $\Delta = 5$, one-block horizon	28.485	87.28	1345.5	4.5	0.193	0	611
4c	MPC margin 80 °C, $\Delta = 5$	31.035	85.15	1310.7	39.3	0.153	0	672
4d	Override alone, $\Delta = 10$, $K_{ov} = 10$	28.530	87.22	1344.2	5.8	0.192	0	—
Limited supply, 1700 A rating								
1	PI and override, $\Delta = 80$ (shipped)	30.146	85.74	1310.3	39.7	0.165	12.28	—
2	MPC under the override, $\Delta = 80$ (scenario 11)	30.522	85.54	1310.3	39.7	0.160	0	656
2′	MPC with the band narrowed to $\Delta = 5$	29.739	86.38	1345.3	4.7	0.174	0	643
3b	Override retuned, $\Delta = 20$, $K_{ov} = 4$	29.607	86.40	1333.8	16.2	0.175	17.02	—
4e	MPC, $\Delta = 80$, without the current row	30.222	85.67	1310.3	39.7	0.164	9.79	638

same diameter, the deposit is the same 118.95 kg in all three and the band trades batch time against margin alone. The default and the wide band never put the center above the limit.

2.6.5 The Predictive Controller Under the Override

Scenario 11 runs the rod of scenario 7 (100 mm, 1100 °C, 2 °C pyrometer noise) under the controller of Section 2.5.4, with the schedule on and the override above it at its shipped band of 80 °C. Every row of Table 2.6 uses the same rod, seed, noise and instruments and runs to harvest. The comparison is a negative result. Under the shipped protection the MPC harvests in 29.940 h against 29.595 h for the override alone, 0.345 h slower, for 85.76 against 85.99 kW h/kg. Its first plan predicts 28.69 h to harvest with the center at 1345.0 °C, but the override does not know the plan is there and takes its own correction off the planned setpoint, so the center settles between 1296.7 and 1299.1 °C from 10 h on and the two margins add. At harvest the plan requests 1059.1 °C and the override applies 1031.4 °C. Without its center row the plan rides at the operator's setpoint and the run is the override's own to the last digit, 29.595 h.

Narrowing the band to 5 °C, so that the override is only a backstop behind the plan, gives the result the controller was built for. The MPC then harvests in 28.482 h with the center held at 1343.2 to 1345.6 °C from 10 h on, 1.11 h sooner than the override alone, for 1.5 % more energy per kilogram and 0.193 against 0.172 ppba of boron, as shown in Figure 2.5. The counterfactuals of Table 2.6 show that the hour is margin given up and not prediction. A one-block plan that looks no further than its next 0.048 h takes 28.485 h, because the center limit restricts only the present setpoint at the present diameter and the optimal profile rides that path constraint [20]. Held to a margin of 80 °C behind the same backstop, the plan takes 31.035 h. A proportional override with a high gain reproduces the constant margin without any model of the future. At $K_{ov} = 10$ and a 10 °C band it harvests in 28.530 h with a least margin of 5.8 °C against the plan's 4.4, and over least margins from 4 to 26 °C the plan and that override lie within 0.016 to 0.031 h of each other at equal margin. Of the 1.11 h, 0.55 h comes from narrowing the margin from 27 to 5 °C (the plan at a margin of 27 behind the backstop takes 29.036 h) and 0.56 h from holding the margin constant rather than letting it grow, which is what the high-gain override also does. The override retuned once by the plan's own rule, to the band at which its steady margin at harvest is the plan's 5 °C, recovers 0.64 and 0.92 h of the gain at gains of 1 and 4. The shipped and narrowed results stand without pyrometer noise (29.594, 29.939 and 28.479 h) and at half the integration step (identical at the printed digits).

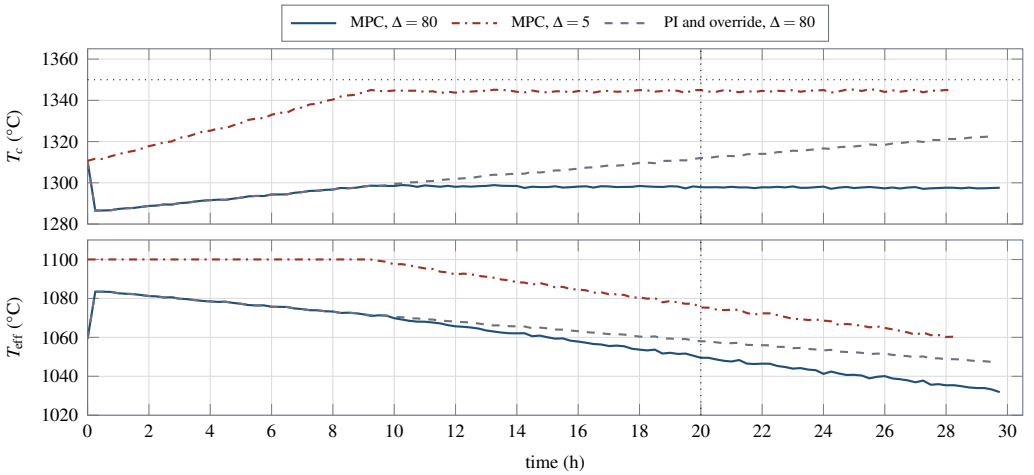


Figure 2.5 True center temperature and effective surface setpoint of the rod of scenario 7 under scenario 11's MPC (solid), the same MPC with the override band at 5 °C (dash-dotted), and the override alone (dashed), sampled from the engine every 0.25 h. At $t = 20$ h (dotted) the MPC under the shipped band holds the center at 1297.9 °C on an effective setpoint of 1049.5 °C while its plan requests 1077.4 °C, the narrowed MPC holds 1345.0 °C on 1075.3 °C, and the override alone holds 1312.0 °C on 1058.0 °C.

Table 2.7 Three Model Errors on the Rod of Scenario 7. Error Is the Metered Center Estimate Less the True Center From 100 to 150 mm. Each Run Gives the Hours Over the Limit and the Peak True Center.

Error	Estimate (°C)	Override $\Delta = 80$	MPC $\Delta = 5$	MPC $\Delta = 80$
$k_{Si} = 18$	−21 to −30	9.57, 1363.8	29.21, 1375.8	0.00, 1356.9
$B + 10\,\%$	−134 to −139	17.28, 1411.5	melts at 27.3	17.28, 1411.5
$L = 6.0$ m	−5 to −11	0, 1326.1	15.33, 1353.8	0, 1310.4

A model error spends that margin. With the band at 5 °C the metered estimate of (2.24) is the only protection left, and Table 2.7 runs it against three parameter changes the controller is not told about. With $k_{Si} = 18$ the estimate reads 21 to 30 °C low and the narrowed MPC keeps the center past the limit for 29.21 h, against 9.57 h for the override alone. With B raised 10 % it reads 134 to 139 °C low and the rod melts at 27.3 h, while the override alone peaks at 1411.5 °C, 2.5 °C short of melting. With rods 6.0 m long it reads 5 to 11 °C low and the center spends 15.33 h past the limit, against none. The shipped band prevents all three. Under it the MPC is never longer over the limit than the override alone (0.005 h in the conductivity case, all of it in the first integration step from the starting state, and none in the rod-length case), and in the resistivity case the override trims the plan to exactly its own run. The shipped band therefore removes both the danger and the gain, and it is the reason scenario 11 keeps it. Under the shipped protection the MPC is slower than the override alone, and a narrower override band gives the same batch time without the MPC.

With the supply rated at 1700 A, just under the 1749 A the shipped run draws two thirds of the way from 100 to 150 mm, the MPC under the shipped band takes 30.522 h and never reaches the rating, against 30.146 h with 12.28 h on the rating for the override alone. The current row is the difference. Without it the MPC takes 30.222 h with 9.79 h at the rating, so the 2 % headroom costs 0.30 h. On the challenge at its suggested settings the MPC takes 30.786 against 29.682 h.

The worst case of every limit was run under each shipped structure and under the MPC, as loaded and with the settings of scenario 11, across the existing scenarios, the challenge and eighteen extreme

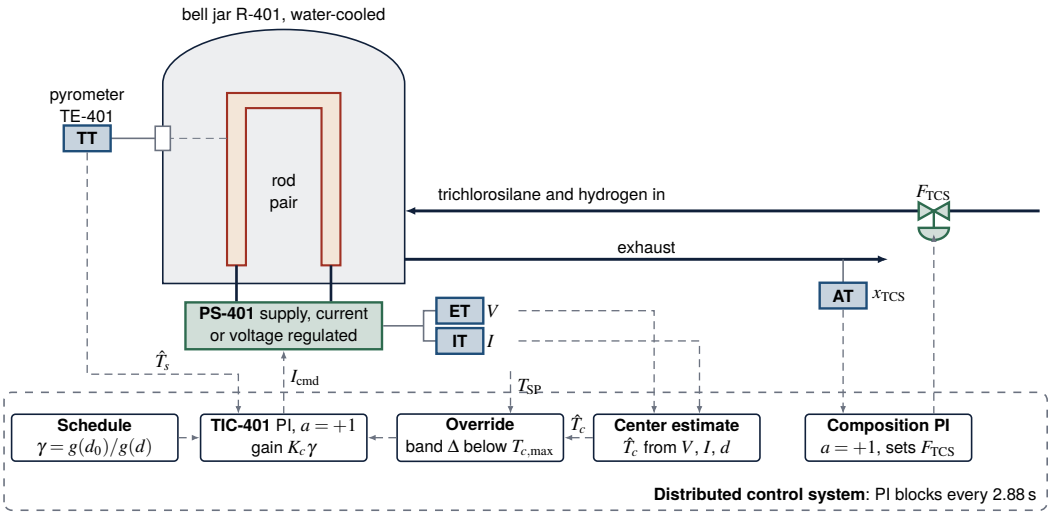


Figure 2.6 The loops of reactor R-401 drawn with their hardware. Pyrometer TE-401 sights a rod through a sight glass, and TIC-401 sets the current command of supply PS-401 with its gain multiplied by the schedule $\gamma(d)$. The supply's own voltmeter and ammeter feed the center estimate, which drives the override that trims the surface setpoint. The composition loop reads the trichlorosilane fraction of the jar gas and sets the feed. Both PI loops are reverse acting ($a = +1$). Notice that the pyrometer enters only TIC-401: the limit that ends the batch is guarded by the electrical meters alone.

settings. Of the 60 MPC runs, 2 are less safe than the better shipped run, 11 are safer and 47 are the same. Neither flag is the plan's doing. On the challenge as loaded the center peaks at 1396.2 against 1395.2 °C at $t = 0.0004$ h, in the loaded state, while the MPC then spends 2.43 h over the limit against 6.93 h and a melt for the shipped structure. Voltage regulation on the 60 mm rod with the schedule of scenario 11 melts at 0.10 h, exactly as the scheduled PI with the same override does.

2.7 Technology in Practice

The loops of Section 2.5 run on a pyrometer, a power supply with its own meters, and a control system, drawn together in Figure 2.6.

What Measures It

The only temperature instrument is pyrometer TE-401. The model forms its reading from the true surface temperature with 2 °C of noise, a fixed offset that a fault can set, and a first-order lag of 1.8 s. A wrong emissivity becomes the bias of Section 1.4.2.

> **TECHNOLOGY** **Sensor: A Radiation Pyrometer Through a Sight Glass**
>
> A radiation pyrometer concentrates the infrared light a hot surface emits onto a small sensing element, often a thermopile, through a lens or a mirror. The emitted power grows as the fourth power of the absolute temperature, a nonlinearity that confines good accuracy to a narrow range [24]. The instrument reads only the surface: a hotter region inside that does not show at the surface is invisible to it [24]. Its main weakness is inaccuracy. The emittance of the target has to be found once against a known temperature, and radiation reflected from other hot bodies taints the reading [24], and a bell jar holds many glowing rods. On a reactor it sights a single spot of one rod [25]. A film on the sight glass makes it read low and steady, the fault of Section 2.6.3.

The second set of measurements is electrical. The supply's voltmeter and ammeter give the rod resistance V/I and the power VI, from which the override infers the mean and center temperatures,

and the deposition rate integrated over the batch gives the diameter. A design for a twelve-rod reactor uses the same two ideas, estimating the diameter from the resistance and the temperature law of silicon and using the resistance as a measurement in its own right [25]. Purity is found only by analyzing a piece of the rod after the batch.

What Moves It

The final element of the temperature loop is supply PS-401. In the model it follows its current or voltage command through a first-order lag of 0.002 h and rests on ratings of 3000 A, 1800 V and 550 kW, with no rectifier dynamics. A rod reading below its setpoint calls for more current, so TIC-401 is reverse acting, $a = +1$ in the convention of Section 1.6.2.

> **TECHNOLOGY** **Final Element: The Rod Power Supply**
>
> A cold silicon rod barely conducts. The reactor of the original patent was heated up at high voltage from transformer taps that held the current to two amperes until the rods glowed red, and then switched to a winding rated for low voltage and high current, on which the current was raised as the rods grew [14]. A newer design rectifies the mains and drives the rods from a resonant transistor inverter above 50 kHz. At that frequency the skin effect moves most of the current toward the outside of the rod, which flattens the radial temperature profile [19]. A supply that regulates the wrong terminal shows on a trend as a current climbing on its own at a steady voltage, the runaway of Section 2.6.2.

The second manipulated variable, the trichlorosilane feed, enters through a vaporizer whose calibration error f_v the model carries as a disturbance.

Where the Controller Runs

Loops of this kind run in a distributed control system, or in programmable controllers acting as its local controllers, with cycle times from milliseconds to minutes [26]. The app executes both PI loops every 2.88 s against a physics step of 1.44 s, both far shorter than the shortest rod time constant of 7.9 s. In a control system the structure of this chapter becomes standard parts: a PID block whose gain is multiplied by the schedule, a calculation block that turns V, I and d into the center estimate, and the override as a trim on the setpoint ahead of the PID. The predictive controller of Section 2.5.4, which decides every 0.048 h, belongs to the supervisory level that sets the setpoints of the loops below it [26].

> **IN PRACTICE** **Operators Lower the Surface as the Rods Grow**
>
> A model study of a twelve-rod reactor found what this chapter finds: over a campaign the time constant of the temperature loop grows by an order of magnitude while its gain falls by more than two, enough to call for adaptive control [25]. Holding the surface constant needs ever more current, which the stresses in the rod and the melting of its core limit, so operators lower the surface temperature as the campaign goes on [25]. Because the pyrometers see single spots, the rod resistance itself can serve as the controlled variable, since it responds even to local temperature changes [25]. The override of Section 2.5.3 automates the operator's retreat.

What Is Modernizing It

Three developments bear on this reactor. The first is adaptive control on electrical estimates: a 2015 design estimates the diameter and temperature from the rod resistance and tunes its loop with neuro-fuzzy methods [25]. The second is nonlinear model predictive control on a lumped model of the reactor, shown on a laboratory-scale Siemens reactor [27]. The shrinking-horizon controller of this chapter belongs to that family. The third changes the plant. A supply above 50 kHz moves the current outward by the skin effect and narrows the difference between center and surface [19], and polished or coated walls cut the radiant loss [8]. Either would let a rod grow further before its center reaches the limit of Section 2.4.3.

2.8 Limitations

The model is not a reactor design tool. A single rod pair with rod-to-wall radiation cannot predict the rod-to-rod uniformity that decides real yield, and the studies that do address it solve a radiation field over tens of rods [4, 12, 13]. The parabolic profile of (2.8) is the load-bearing approximation of the whole application, and it rests on uniform generation and constant conductivity, neither of which holds in hot silicon. The chemistry is a single overall stoichiometry with no dichlorosilane, no powder formation and no hydrochlorination recycle, so nothing in the application can speak to the conversion efficiency or the byproduct economics that dominate a real plant [6, 10]. The cold start is absent entirely, which removes the most awkward control problem the process has. The supply is two algebraic expressions and a first-order lag, with a current rating, a voltage rating and a power ceiling but no rectifier dynamics, no transformer tap and no minimum firing angle, so the transition from voltage regulation into the power ceiling and the current limit in the counterfactual run of Section 2.6.2 is instantaneous where a real supply would take a cycle or two. Disturbances arrive as commanded steps and never as the drifting behavior of a real plant. Repeatability is itself unrealistic, because a fixed integration step and a seeded generator reproduce one trajectory exactly while a distributed control system does no such thing. The three features that survived the simplification, the 785-fold change in resistance, the sign of (2.16) under the two regulation modes and the linear growth of the center-to-surface rise, survived because a controller has to be designed around each of them. Where the individual assumptions break is set out in Section 2.3.6. The predictive controller's model is the quasi-steady design point at the declared parameters. It leaves out the transients of the surface and composition loops, it knows nothing of a learner's conductivity, resistivity law or rod length, and its biases cannot absorb an error in the center estimate itself. It does not model the override above it, so its predicted batch time and center describe the plan rather than the trimmed run, and a plan that did model the override could at best reproduce the override's own run, which is what the plan without its center row does. Three linearization passes are not always enough. A decision taken 140 °C above its answer, at a setpoint of 1200 °C on a 150 mm rod, was accepted with 34 °C of slack and the next decision was exact, a case that no scenario reaches.

> ⚠ **OPEN CHALLENGE** **A Core That No Instrument Reaches**
>
> The temperature gradient inside the rods remains the main limit of rod deposition. Silicon conducts better as it heats, so the current crowds toward the hot center and heats it further, and the gradient stresses the rod and limits how large it can grow before its core melts [19]. The center is still estimated numerically, and the gradients inside the rods are not controlled [25]. Progress would be a center estimate that is checked against the plant during a batch, or a heating current that keeps the profile flat enough for the estimate to matter less.

2.9 Discussion

The energy number on the scorecard is not decoration. A complete batch under the scheduled loop with the override in service consumes 18 564 kW h for 213.5 kg of silicon, or 87.0 kW h/kg, and polysilicon production is a substantial term in the life-cycle energy of a photovoltaic module [11]. Both of the levers that improve it are constrained by something other than the loop. Running hotter raises the rate by 8.3 % for sixty degrees and costs 39 % more boron, and running to a larger diameter improves the energy per kilogram until the center constraint stops it, which is the trade the deposition-radius optimization literature studies directly [5, 13]. The application exposes the harvest diameter, the core limit and the thermal conductivity as adjustable parameters so that the trade can be computed rather than asserted. Moving k_{Si} alone moves the diameter at which the batch has to stop, which is an unusually direct demonstration that a material property can be the binding constraint on a control problem. The override infers the center at the declared 22 W/(m K), so a different k_{Si} is a model error it cannot see: at 18 its center reads up to 30 °C low and the true center spends 9.6 h up to 14 °C past the limit.

The model of this chapter also supports methods beyond the implemented loops. The predictive controller of Section 2.5.4 already treats the surface-temperature trajectory as the decision variable with the constraint of (2.8) enforced explicitly, and its result says where the remaining value lies. A constant margin is something a high-gain override already provides, so what a model could add is a margin sized to its own uncertainty, which needs an estimate that reports how wrong it may be. The estimator in (2.8) is the simplest member of a family, and moving horizon estimation on the same equations would reconstruct the loss multiplier and the pyrometer bias from the current and the composition, which is what would turn the cross-check of Section 2.6.3 into a diagnosis. A purity specification that is violated with every instrument nominal is the standing argument for soft sensing on the feed side rather than on the reactor side [17].

2.10 Conclusion

This chapter studied a Siemens polysilicon reactor as an interactive laboratory for plants that are non-stationary by construction. The simulator integrates nine states covering the rod diameter and mean temperature, two supply lags, the trichlorosilane and tetrachloride mole fractions, the deposited silicon, the incorporated boron and the consumed energy, with a semiconducting resistivity, a radiative and convective loss on the rod surface and a parabolic radial profile that places the hottest point on the axis. Growing an 8 mm seed to 150 mm takes the resistance from 50.9 to 0.0649 Ω, the current from 21.8 to 2646 A and the steady-state gain from 10.46 to 0.0828 °C/A, a factor of 126 with nothing disturbed. A gain schedule built from the same closed-form expressions cuts the batch integrated error from 760.4 to 43.1 °C h and holds the surface inside 2.24 °C over runs of 80.14 and 78.57 h, and it melts the rod at 149.3 mm because the constraint that ends a batch is at the center and not at the surface. Adding an override on that inferred center caps it at 1323 °C and harvests 213.5 kg at 87.0 kW h/kg. The same term $PB/\bar{T}^2$ that makes constant-voltage operation unstable by a factor between 2.03 and 2.50 against the loss slope is also the larger part of the loop-gain denominator, and holding the terminal voltage of a 60 mm rod near 346 V carries its center to the melting point in 6.50 min. A shrinking-horizon predictive controller of the surface setpoint takes 29.94 h against 29.60 h under the shipped override, and the 1.11 h it saves with the band narrowed to 5 °C is margin that a high-gain override recovers to within 0.03 h and that a 10 % error in the resistivity law spends to a melt.

A fuller model would add rod-to-rod view factors, a temperature-dependent conductivity, a cold-start sequence and a rectifier model for the supply. As it stands, the reactor gives practice in controlling a plant that is also the product. The Czochralski puller of Chapter 3, the next step of the wafer flow, also grows its product while it runs and also meets a limit that no instrument reports.

KEY POINTS

- The silicon rod is heater, deposition surface and product at once, so the plant changes while it runs.
- Growing from 8 to 150 mm takes the plant gain from 10.46 to 0.0828 °C/A, a factor of 126.
- Holding the current makes the rod self-regulating, while holding the voltage carries the center of a 60 mm rod to the melting point in 6.50 min.
- A schedule built from the closed-form gain cuts the batch IAE from 760.4 to 43.1 °C h but melts the rod, because the limit is at the unmeasured center.
- An override on the center inferred from the supply's meters holds it to 1323 °C and harvests 213.5 kg.

Where else this applies. An exact gain schedule handed over by the physics in Chapter 3, and a loop that holds its pyrometer reading while the product is wrong in Chapter 9.

References

[1] J. S. Shamma and M. Athans, "Analysis of gain scheduled control for nonlinear plants," *IEEE Transactions on Automatic Control*, vol. 35, no. 8, pp. 898–907, 1990.

[2] W. J. Rugh and J. S. Shamma, "Research on gain scheduling," *Automatica*, vol. 36, no. 10, pp. 1401–1425, 2000.

[3] D. J. Leith and W. E. Leithead, "Survey of gain-scheduling analysis and design," *International Journal of Control*, vol. 73, no. 11, pp. 1001–1025, 2000.

[4] G. del Coso, I. Tobías, C. Cañizo, and A. Luque, "Temperature homogeneity of polysilicon rods in a Siemens reactor," *Journal of Crystal Growth*, vol. 299, no. 1, pp. 165–170, 2007.

[5] A. Ramos, W. O. Filtvedt, D. Lindholm, P. A. Ramachandran, A. Rodríguez, and C. del Cañizo, "Deposition reactors for solar grade silicon: A comparative thermal analysis of a Siemens reactor and a fluidized bed reactor," *Journal of Crystal Growth*, vol. 431, pp. 1–9, 2015.

[6] B. Ceccaroli and O. Lohne, "Solar grade silicon feedstock," in *Handbook of Photovoltaic Science and Engineering*, 2nd ed., A. Luque and S. Hegedus, Eds. Chichester, UK: John Wiley & Sons, 2010, pp. 169–217.

[7] A. F. B. Braga, S. P. Moreira, P. R. Zampieri, J. M. G. Bacchin, and P. R. Mei, "New processes for the production of solar-grade polycrystalline silicon: A review," *Solar Energy Materials and Solar Cells*, vol. 92, no. 4, pp. 418–424, 2008.

[8] K. Shen, W. Jin, and J. Wang, "Numerical simulation of heat transfer process and heat loss analysis in Siemens CVD reduction furnaces," *Frontiers in Heat and Mass Transfer*, vol. 22, no. 5, pp. 1361–1379, 2024.

[9] G. Bye and B. Ceccaroli, "Solar grade silicon: Technology status and industrial trends," *Solar Energy Materials and Solar Cells*, vol. 130, pp. 634–646, 2014.

[10] W. C. Breneman and S. Julsrud, "From conventional polysilicon Siemens process to low-energy fluidized bed processes using silane," in *Solar Silicon Processes*, B. Ceccaroli, E. Ovrelid, and S. Pizzini, Eds. Boca Raton, FL: CRC Press, 2016, pp. 199–230.

[11] V. M. Fthenakis and H. C. Kim, "Photovoltaics: Life-cycle analyses," *Solar Energy*, vol. 85, no. 8, pp. 1609–1628, 2011.

[12] G. del Coso, C. del Cañizo, and A. Luque, "Radiative energy loss in a polysilicon CVD reactor," *Solar Energy Materials and Solar Cells*, vol. 95, no. 4, pp. 1042–1049, 2011.

[13] Z. Nie, P. A. Ramachandran, and Y. Hou, "Optimization of effective parameters on Siemens reactor to achieve potential maximum deposition radius: An energy consumption analysis and numerical simulation," *International Journal of Heat and Mass Transfer*, vol. 117, pp. 1083–1098, 2018.

[14] H. Schweickert, K. Reuschel, and H. Gutsche, "Production of high-purity semiconductor materials for electrical purposes," U.S. Patent 3,011,877, filed 11 June 1957 (priority Germany, 25 June 1956), 1961, granted 5 December 1961.

[15] H. Habuka, T. Nagoya, M. Mayusumi, M. Katayama, M. Shimada, and K. Okuyama, "Model on transport phenomena and epitaxial growth of silicon thin film in $SiHCl_3$–H_2 system under atmospheric pressure," *Journal of Crystal Growth*, vol. 169, no. 1, pp. 61–72, 1996.

[16] B. Joseph and C. B. Brosilow, "Inferential control of processes: Part I. Steady state analysis and design," *AIChE Journal*, vol. 24, no. 3, pp. 485–492, 1978.

[17] P. Kadlec, B. Gabrys, and S. Strandt, "Data-driven soft sensors in the process industry," *Computers & Chemical Engineering*, vol. 33, no. 4, pp. 795–814, 2009.

[18] D. Krishnamoorthy and S. Skogestad, "Systematic design of active constraint switching using selectors," *Computers & Chemical Engineering*, vol. 143, p. 107106, 2020.

[19] K. A. Kozin, A. G. Goryunov, and F. Manenti, "Siemens-reactor's high-frequency power supply," *Chemical Engineering Transactions*, vol. 39, pp. 1651–1656, 2014.

[20] B. Srinivasan, S. Palanki, and D. Bonvin, "Dynamic optimization of batch processes," *Computers & Chemical Engineering*, vol. 27, no. 1, pp. 1–26, 2003.

[21] M. Ellis, H. Durand, and P. D. Christofides, "A tutorial review of economic model predictive control methods," *Journal of Process Control*, vol. 24, no. 8, pp. 1156–1178, 2014.

[22] G. del Coso, C. del Cañizo, I. Tobías, and A. Luque, "Increase on Siemens reactor throughput by tailoring temperature profile of polysilicon rods," in *2007 Spanish Conference on Electron Devices*. Madrid, Spain: IEEE, 2007, pp. 25–28.

[23] D. Goldfarb and A. Idnani, "A numerically stable dual method for solving strictly convex quadratic programs," *Mathematical Programming*, vol. 27, no. 1, pp. 1–33, 1983.

[24] T. R. Kuphaldt, *Lessons in Industrial Instrumentation*, version 2.33 ed. Open textbook, Creative Commons Attribution 4.0 license, ibiblio.org/kuphaldt/socratic/sinst, 2024.

[25] K. Kozin, A. Goryunov, F. Manenti, F. Rossi, and A. E. Stolpovskiy, "Development of an advanced control system for a chemical vapor deposition (CVD) reactor for polysilicon production," *Chemical Engineering Transactions*, vol. 43, pp. 1531–1536, 2015.

[26] K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, A. Hahn, S. Saravia, A. Sherule, and M. Thompson, "Guide to operational technology (OT) security," National Institute of Standards and Technology, Gaithersburg, MD, Special Publication 800-82r3, 2023.

[27] L. Viganò, M. Vallerio, F. Manenti, N. M. N. Lima, L. Zuniga Linan, and G. Manenti, "Model predictive control of a CVD reactor for production of polysilicon rods," *Chemical Engineering Transactions*, vol. 21, pp. 523–528, 2010.

2.11 Exercises

2.1 Concept check · 2 min

A 60 mm rod sits at its design point with the surface at 1100 °C and no controller acting. The supply is switched from current regulation to voltage regulation at the same terminal voltage. What does the model predict?

(a) The rod holds its temperature, because the operating point is an equilibrium in either mode

(b) The rod cools slowly, because a fixed voltage delivers less power than a fixed current

(c) The rod oscillates, because the supply lag and the rod time constant now interact

(d) The rod runs away, because the power then rises with temperature faster than the loss

2.2 Concept check · 2 min

The schedule of Equation (2.18) is built from $2IR/S$, which overstates the measured plant gain $\partial T_s/\partial I$ by a factor of 3.449 at the seed and 3.593 at harvest. Why does the schedule still give the loop the same shape at every diameter?

(a) The factor is nearly constant, so the gain entered at the seed absorbs it once

(b) The factor is large, so the scheduled loop is detuned and robust at every diameter

(c) The integral action removes the factor, because a PI loop leaves no steady error

(d) The factor has no effect, because the multiplier is clamped between 0.05 and 400

2.3 Concept check · 1 min

In scenario 10 the boron in the feed triples at $t = 2\,\mathrm{h}$. Which signal in the reactor responds?

(a) The pyrometer reading, because boron changes the emissivity of the rod surface

(b) None of them, because temperature, current, composition and growth do not depend on boron

(c) The rod current, because a doped deposit conducts better and lowers the rod resistance

(d) The gas composition, because boron displaces trichlorosilane in the feed

2.4 Theory and calculation · 6 min

Use Equations (2.1) and (2.2) with $\rho_\infty = 2.0 \times 10^{-5}\,\Omega\,\mathrm{m}$, $A = 7.32 \times 10^{-8}\,\Omega\,\mathrm{m}$, $B = 12\,120\,\mathrm{K}$ and $L = 5.2\,\mathrm{m}$. With the surface held at 1100 °C, the mean rod temperature is 1108.43 °C on the 8 mm seed and 1258.03 °C at 150 mm.

(a) Find the resistivity and the resistance of the rod pair on the seed. *Checkpoint:* $4.925 \times 10^{-4}\,\Omega\,\mathrm{m}$ and $50.95\,\Omega$.

(b) The seed dissipates 24.23 kW. Find the current and the terminal voltage. *Checkpoint:* 21.81 A and 1111 V.

(c) Find the resistance at 150 mm, the ratio to the seed value, and how much of that ratio is area and how much is resistivity. *Checkpoint:* $0.0649\,\Omega$, a ratio of 785, of which 351.6 is area and 2.23 is resistivity.

Give each answer to within 0.5 %.

2.5 Theory and calculation · 6 min

A 60 mm rod pair of length $L = 5.2\,\text{m}$ has its surface at 1100 °C, a mean temperature of 1163.2 °C and a dissipation of 181.7 kW. Take $\varepsilon = 0.70$, $\sigma = 5.670 \times 10^{-8}\,\text{W}/(\text{m}^2\,\text{K}^4)$, $h_g = 60\,\text{W}/(\text{m}^2\,\text{K})$, $f_{\text{loss}} = 1$, $B = 12\,120\,\text{K}$, $\rho_{Si} = 2330\,\text{kg}/\text{m}^3$ and $c_{p,Si} = 800\,\text{J}/(\text{kg}\,\text{K})$.

(a) Find the loss slope S of Equation (2.7). *Checkpoint:* 461.7 W/K.

(b) Find $PB/\bar{T}^2$, its ratio to S, and say which regulation mode holds the operating point. *Checkpoint:* 1067 W/K, 2.31 times S. Current regulation.

(c) Find the rod thermal time constant $\rho_{Si}\pi d^2 L c_{p,Si}/(4S)$. *Checkpoint:* 59.4 s.

2.6 Theory and calculation · 6 min

With the surface held at 1100 °C the dissipation is proportional to the diameter, 454.4 kW at 150 mm. Take $k_{Si} = 22\,\text{W}/(\text{m}\,\text{K})$ and $L = 5.2\,\text{m}$.

(a) Find the center-to-surface rise $\Delta T = P/(4\pi k_{Si} L)$ at 150 mm. *Checkpoint:* 316.1 °C.

(b) Find the diameters at which the center reaches the 1350 °C limit and the 1414 °C melting point. *Checkpoint:* 118.6 mm and 149.0 mm, each within 0.2 mm.

(c) The rod's true conductivity is 18 W/(m K), but the estimator of Equation (2.24) assumes 22. Find by how much the estimate of T_c reads low per 100 kW of measured power, and the power range that the error of −21 to −30 °C in Table 2.7 implies. *Checkpoint:* 7.73 °C per 100 kW. About 270 to 390 kW.

2.7 App challenge · 8 min

Load scenario 3, The same tuning at 140 mm, and change no setting. The setpoint steps from 1050 to 1070 °C at 0.5 h. Run with the default seed at the default speed of 0.5 h/s, pause within 0.05 h of $t = 4.0\,\text{h}$ and export the CSV file. Then do the same with scenario 4, Scheduling on a diameter the tool knows exactly.

(a) In the scenario 3 file, read the true surface, column `ts_C`, at $t = 3.0\,\text{h}$, and read the IAE tile of the scorecard. *Checkpoint:* 1053.2 °C, within 0.1 °C. IAE 65.0 °C h, within 1 °C h.

(b) In scenario 4, read the IAE and Overshoot tiles, and the schedule multiplier, column `sched_gain_x`, at $t = 1.0\,\text{h}$. *Checkpoint:* IAE 5.9 °C h, within 0.2 °C h. Overshoot 15.6 %. Multiplier 51.2, within 0.1.

(c) The true overshoot of scenario 4 is 0.40 °C, or 2.0 % of the step. Explain why the tile reads more.

2.8 App challenge · 8 min

Load scenario 2, The loop closed, on a small rod. Before Run, set Gain scheduling on diameter to On. Run with the default seed at 10 h/s until the run stops itself, which takes about eight seconds.

(a) Read the status line, the Diameter tile and the IAE tile of the scorecard. *Checkpoint:* The rod melted, at 149.3 mm. IAE 43.1 °C h, within 0.2 °C h.

(b) Press Reset, set Core temperature override to On, and run again until the run stops. Read the Diameter, Silicon and Energy tiles, the IAE tile, and the largest value of column `t_core_C` in the exported file. *Checkpoint:* 150.0 mm, 213.5 kg, 87 kW h/kg, IAE 38.8 °C h, peak center 1322.7 °C within 0.5 °C.

(c) Compare both runs with the PI-only column of Table 2.5, and explain why the better loop of part (a) lost the rod.

2.9 App challenge · 9 min

Load the graded challenge, Challenge · finish the batch without melting it. Before Run, set Gain scheduling on diameter to On, Core temperature override to On and Rod surface temperature to 1100 °C. Run with the default seed at 10 h/s until the run stops at harvest, and export the CSV file. Skip its comment lines, which start with #. Repeat with Override band below the core limit (in Advanced) at 40 °C and then at 120 °C, pressing Reset before each run.

(a) For each band, read the harvest time from the clock and, from the rows with `time_h` of 1 or more, the largest value of `t_core_C`. *Checkpoint:* Band 80: 29.68 h, 1348.8 °C. Band 40: 28.86 h, 1374.5 °C. Band 120: 30.62 h, 1323.1 °C. Times within 0.02 h, temperatures within 0.3 °C.

(b) Say which runs pass the challenge, and find the batch time each band trades for each degree of margin. *Checkpoint:* Bands 80 and 120 pass. About 0.032 h per degree from 40 to 80, and 0.036 h per degree from 80 to 120.

2.10 Application area · 3 min

An engineer proposes raising the surface setpoint from 1100 to 1160 °C to shorten the batch. Use Equation (2.10) with $k_s = 3.60\,\mathrm{mm/h}$ at 1100 °C, 5.190 mm/h at 1160 °C and $k_m = 1.20\,\mathrm{mm/h}$, at $x_{\mathrm{TCS}} = x_{\mathrm{ref}}$. Use Equation (2.14) with $C_B = 0.60$ ppba, $k_{B0} = 0.35$ and $k_{B1} = 0.55$ per 100 °C.

(a) Find the diameter growth rate at each setpoint. *Checkpoint:* 1.800 and 1.949 mm/h, 8.3 % more.

(b) Find the boron content of the deposit at each, and give a recommendation. *Checkpoint:* 0.210 and 0.292 ppba, 39 % more, against a 0.30 ppba specification.

2.11 Application area · 4 min

Table 2.3 lists 1111 V at 21.8 A on the seed and 172 V at 2646 A at harvest, where the rods draw 454.4 kW. A complete scheduled batch with the override consumes 18 564 kW h for 213.5 kg, and Table 2.2 prices electricity at $0.06 per kW h.

(a) Find the electricity cost of the batch and per kilogram. *Checkpoint:* $1114 per batch, $5.22 per kilogram.

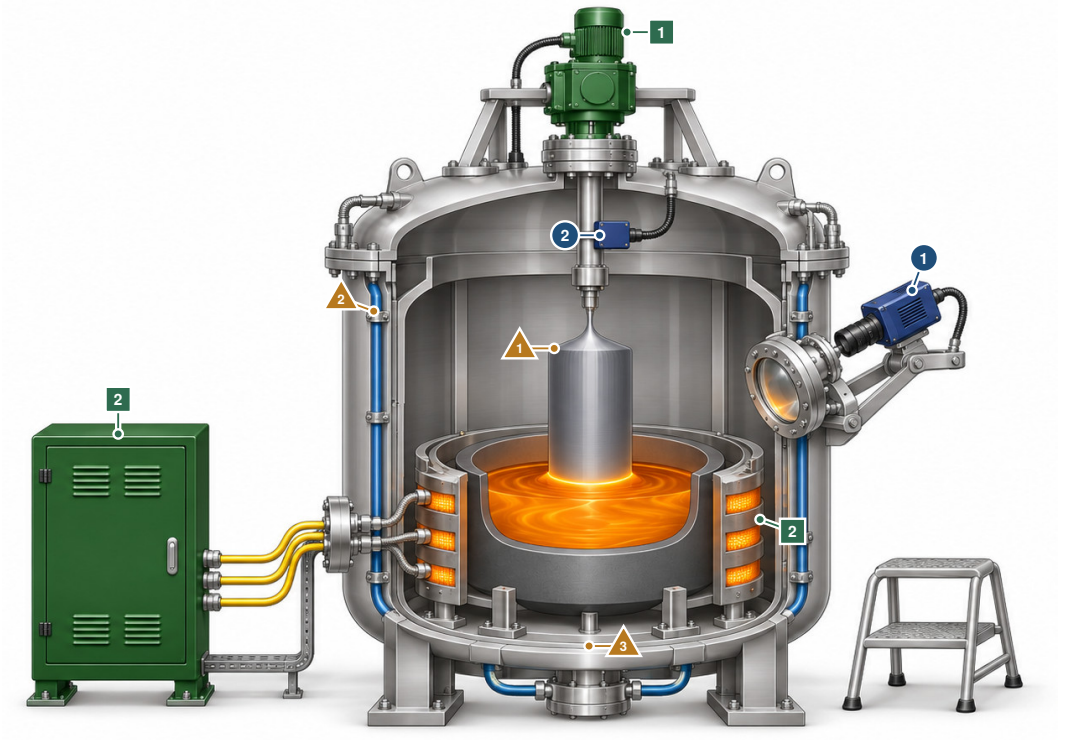
(b) A supply that could deliver the seed voltage and the harvest current at once would be rated at their product. Find that rating, compare it with the largest power, and say what a supply designer does about it. *Checkpoint:* 2.94 MV A, 6.5 times the 454 kW.

2.12 Review the agent · 5 min

An AI assistant reviews scenario 11 and writes: “With the override band narrowed to 5 °C, the MPC harvests in 28.48 h against 29.60 h for the shipped override alone. Its look-ahead saves 1.11 h on every batch. Narrow the band to 5 °C on the plant and run the MPC. The override is still there as a backstop, so the change adds no risk.” Find the flaw and say how you would check it in the app.

CHAPTER 3

Czochralski Crystal Puller Simulator for Mid-Ranging a Fast Limited Actuator



The controller moves 1 pull drive, set by the diameter loop
2 heater around the crucible and its power supply, set by the mid-ranging loop

The controller reads 1 camera on the bright ring, which gives the diameter
2 the pull rate the drive applies, read by the mid-ranging loop

What disturbs it 1 argon flow that changes how fast the crystal freezes
2 a hot zone that loses more heat than the recipe assumed 3 a crucible lift that drifts

Notice the bright ring at the foot of the crystal: the camera reads the diameter there, while the pull-rate ceiling that ends a good crystal has no instrument at all. Illustration, not to scale.

IN PLAIN TERMS **Pulling a Crystal at a Safe Speed**

The main idea. Computer chips are cut from one large crystal of silicon. It is drawn slowly upward out of a pot of melted silicon, and the pulling speed sets how thick it grows, through a short column of liquid that clings to its bottom.

The challenge. Pulling faster makes the crystal thinner, so an automatic control holds its size by changing the pulling speed. Each upset it corrects leaves the pull a little faster. Above a speed limit that no instrument shows, the crystal stops being one perfect crystal, and what grows after is scrap.

Why it matters. A second, slower control on the heater gives the pulling speed back its room and saves the whole crystal.

THREE WAYS TO READ THIS CHAPTER

Explore	11 min	The first page, the fact and history boxes in Sections 3.1 and 3.2, the try-it and fact boxes in Section 3.6, and the key points
Study	60 min	The whole chapter with its pause points and the twelve exercises
Practice	20 min	The first page, Section 3.6, Section 3.7, and Section 3.8 with its open challenge

The control question The pull rate reaches the crystal diameter only through the meniscus, two integrations away, and it has a ceiling that falls as the crystal fattens and that no instrument reports. How is the diameter held without spending that margin, and what does a slow heater loop closed on the pull rate buy?

Learning objectives

- Derive the double-integrator diameter plant and explain why derivative action is the whole of its phase margin.
- Derive the gain schedule on the pull rate that preserves the loop shape, and say what the fixed camera and drive lags leave out of it.
- Size the separation between a mid-ranging heater loop and the diameter loop from the static path from power to pull rate.
- Read the pull-rate margin and the prime crystal, not the diameter, to judge a run, and trade derivative noise gain against filter lag.
- Compare one predictive plan for both actuators with two retuned loops on the same batch.

Prerequisites Sections 1.6.2 (velocity form and loop action), 1.6.3 (the predictive option), 1.6.4 (scenarios, scorecard and CSV) and 1.4.8 (control within a batch). Chapter 2 schedules a gain along a batch.

The app https://apmonitor.com/apps/cz-crystal-growth/

See also Volume 4, Chapter 10 and Volume 4, Chapter 9 for mid-ranging and valve-position control on power plants. Volume 7, Chapter 5 for derivative action on a double integrator. Volume 8, Chapter 4 and Volume 8, Chapter 6 for batches that follow a trajectory instead of a steady state.

3.1 Introduction

The only mechanical connection between a growing silicon crystal and the melt it comes from is a column of liquid a few millimeters tall, held up by surface tension. Everything the pull drive does to the diameter of the crystal it does through that column. Withdrawing the crystal faster than it solidifies lengthens the column, a longer column lowers the angle the solid makes with the free surface at the three-phase junction, and a lower angle narrows the crystal from that moment onward. Two integrations separate the actuator from the controlled variable, so a step in pull rate produces no immediate diameter change whatever and then an accelerating one, and a permanent pull-rate error is not a diameter offset but a cone. A loop built on that plant arrives at its crossover frequency with 180° of lag before the controller has contributed anything.

The second actuator is the heater, and it is the opposite in every respect. Heater power reaches the interface through a graphite element, a quartz crucible and a hundred kilograms of molten silicon, which in the model of this chapter is a dominant time constant of 84.8 min against a diameter loop whose crossover period is 36.3 min. The heater has authority without speed. The pull drive has speed without authority, and it has something worse than a small gain, which is a ceiling: above a pull rate set by how fast the latent heat of solidification can leave the interface up the crystal, the growth front breaks down and every kilogram grown afterward is polycrystalline. That ceiling falls as the crystal fattens and no instrument on the machine reports it.

✦ DID YOU KNOW? From 50 to 100 Kilograms

In 1985 production pullers melted charges of 50 kg of polysilicon, and the largest wafers in production were 150 mm across [1]. As of 2009, more than 95 percent of the 300 mm single-crystal silicon ingots were grown by the Czochralski process [2]. This chapter's puller grows a 300 mm crystal from a 100 kg charge.

A fast actuator with a hard limit paired with a slow actuator that has authority is a standard industrial arrangement, and the standard answer is not to choose between them. The fast loop is closed on the product and a second, slower loop is closed on the first loop's own output, so that the fast actuator is found near the middle of its range when the next disturbance arrives. The structure carries several names, valve position control and mid-ranging among them, and it is documented as a general design rather than a trick [3, 4]. Volume 4, Chapter 10 and Volume 4, Chapter 9 build the same structure on a power cycle and on a trigeneration plant. Readers will recognize the same shape in engine torque against transmission ratio, in a surge valve against a compressor recycle, and in every trim heater sitting behind a control valve. The Czochralski puller is an unusually honest instance of it, because the fast actuator's limit is a process constraint rather than a stop on the drive, and because the growth rate carries a two-sided product specification of its own [5, 6].

Czochralski dynamics and control have their own literature. The diameter response through the meniscus was written down and analyzed in the 1970s [7, 8], a state-space treatment with an explicit thermal model followed [9], and model-based designs with observers and scheduled parameters are current [10, 11]. However, a student handed the transfer function $-2V_p c_h/(s^2+\omega_n^2)$ and asked for a PID setting never meets the part of the problem that decides whether the run survives, which is that holding the diameter perfectly spends an actuator nobody is watching. This chapter sets out an interactive Czochralski puller in which the diameter loop always succeeds and the run is lost anyway, so that the remedy has to be a second loop rather than a better tuning. Section 3.2 traces how diameter control became automatic. Section 3.3 gives every equation and constant of the ten-state puller, with its meniscus, its thermal chain, its draining charge, its optical instrument and its actuators, at the level of detail a rebuild requires. Section 3.4 derives the double integrator, the exact gain schedule that the meniscus equations hand the diameter loop and the phase margin it holds, and the separation the heater path needs. Section 3.5 configures the two loops and the optional predictive controller. Section 3.6 runs the scenario sequence, in which the pull-rate margin, not the diameter, is the number that predicts the outcome. Section 3.7 names the camera, the drives and the controller behind the loops. Sections 3.8 and 3.9 state what the model leaves out and where model-based designs take the problem. The growth model, the mid-ranging structure and the predictive controller are all established methods.

3.2 Automatic Diameter Control

The method is named for a 1918 measurement of crystallization velocity in which a thin filament was drawn from molten metal and its length recorded against time [12]. What made it an industrial process was the demand for germanium and then silicon single crystals for solid-state devices, and by the early 1980s the pulling of silicon had its own detailed process literature covering the charge, the hot zone, the argon sweep and the defect structure of the product [13].

Diameter control was manual for the first three decades. An operator watched the bright ring of reflected light where the meniscus meets the crystal and moved the pull rate or the heater by hand, and the shape of the finished boule was a permanent record of how attentive the operator had been. Two independent routes to an automatic measurement appeared in the same volume of proceedings in 1972. One is optical, using a television camera aimed at the bright ring and an electronic threshold to convert its apparent width into a diameter signal [14]. The other weighs the crystal, or the crucible, and differentiates the load cell signal to recover the rate at which mass is being solidified [15]. The weighing route required its own theory before it was trustworthy, because the load cell reads the weight of the meniscus as well as the weight of the crystal and the two contributions move in opposite

directions during a diameter transient [16]. The application of this chapter takes the optical route and carries the camera filter and camera noise explicitly, because those two numbers set what derivative action costs.

Measurement was only half the problem. A controller needs a model, and the model that makes the diameter loop tractable is the dynamic meniscus, in which the growth angle is a function of the height of the melt column and the radius integrates the departure of that angle from its equilibrium value [7, 8]. That relation is what turns the plant into a double integrator and what forces a derivative time comparable to the closed-loop settling time. The two-actuator structure followed from furnace economics. As charge weights grew from a few kilograms to a hundred and more, a heater correction large enough to matter took longer than the diameter excursion it was meant to fix, so the fast correction moved permanently to the pull drive and the heater became the trim. Later work made the split explicit, treating the choice of manipulated variable and the scheduling of controller parameters along the batch as design questions in their own right [9, 11].

> ⧖ HISTORY **A Heater Loop That Waits and Sees (1972)**
>
> A paper received in October 1972 reported automatic diameter control of a silicon puller [17]. Wafers were moving from two to three inches, crystals free of dislocations had become the norm, and holding the diameter by hand had grown difficult. Optical methods already sighted the meniscus with a radiation pyrometer or a television camera. The paper weighed the crystal instead and moved the pull rate, with strong derivative action to keep that loop from oscillating. Control through the heater alone had not succeeded: its response was slower, its dead time about twice as long, and its gain too sensitive to temperature to measure [17]. So a second controller trimmed the heater power whenever the pull rate left a set band, acting slowly in a “wait and see” manner so that the two would not oscillate [17]. It is an early form of the mid-ranging structure of this chapter.

3.3 Puller Model

Puller CZ-300 grows one 300 mm silicon crystal from a 100 kg charge. A graphite heater surrounds a quartz crucible that holds the melt, the crucible is lifted as the charge drains so that the melt surface stays at a fixed height, and the seeded crystal is withdrawn through a meniscus of height h at pull rate V_p. Bright-ring camera DT-301 reports the diameter, diameter controller DIC-301 drives the pull rate, and mid-ranging controller VIC-301 reads the pull rate and drives the heater supply. Figure 3.1 shows the arrangement and every symbol used in the equations that follow. The application starts in body growth and the seed, neck, crown and shoulder are outside its scope.

Table 3.1 lists what the two loops interact with, with D read by DT-301, V_p driven by DIC-301 and P driven by VIC-301. The plant is written in the units of a puller control room, lengths in millimeters, rates in millimeters per minute, mass in kilograms, temperature in degrees Celsius, power in kilowatts and time in minutes. Ten states are integrated.

3.3.1 The Meniscus and the Radius

The melt column between the crystal and the melt surface lengthens whenever the crystal is withdrawn faster than it solidifies, and the crucible lift is supposed to hold the melt surface still while the charge drains. Both effects enter the same integrator, as shown in Equation (3.1).

$$\frac{\mathrm{d}h}{\mathrm{d}t} = V_p - V_g + \varepsilon_L \tag{3.1}$$

where h is the meniscus height in mm, V_p is the pull rate, V_g is the growth rate and ε_L is the crucible lift error, all in mm/min. Silicon grows at a fixed angle to the free surface of the meniscus, so the angle the solid makes with the vertical departs from zero only when the column departs from its equilibrium height, and the radius integrates that departure as shown in Equation (3.2) [7, 8].

$$\frac{\mathrm{d}r}{\mathrm{d}t} = V_p \tan\big[\alpha(h)\big], \qquad \alpha(h) = -c_h\,(h - h_0) \tag{3.2}$$

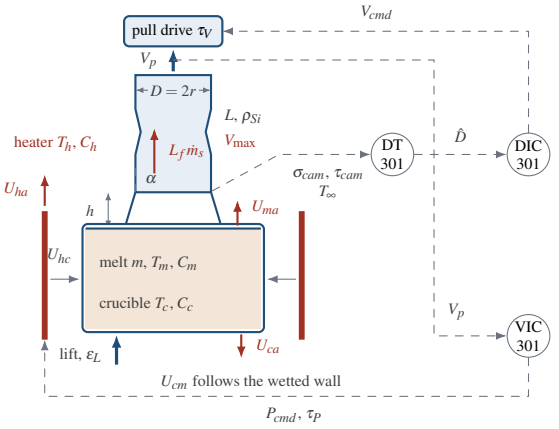


Figure 3.1 Czochralski puller CZ-300 in section, with the melt column that is the only path from pull rate to diameter. The crystal of radius r and length L is withdrawn at V_p through a meniscus of height h, and the growth angle α at the three-phase junction falls as h rises. Latent heat $L_f \dot{m}_s$ leaves the interface up the crystal and sets the pull rate $V_{\max}$ above which structure is lost. Heater power enters node T_h and reaches the melt T_m through the crucible node T_c. Bright-ring camera DT-301 supplies $\hat{D}$ with noise σ_{cam} and filter τ_{cam}, DIC-301 commands the pull drive, and VIC-301 reads the pull rate and commands the heater.

Table 3.1 What the Two Loops Interact With. The Nominal Column Is the Design Point That Opens Every Scenario but the Tail Scenario.

Symbol	Role	Nominal	Range	Unit
$D = 2r$	CV, measured	300.00	270–330 SP	mm
V_p	MV, and PV of VIC-301	1.000	0.05–2.50	mm/min
P	MV of VIC-301	118.26	30–220	kW
V_g	output, measured	1.000	0.72–1.32 spec	mm/min
$V_{\max}$	output, unmeasured	1.600	—	mm/min
η	DV, measured	1.000	0.700–1.300	—
λ	DV, measured	1.000	0.800–1.300	—
ε_L	DV, unmeasured	0	±0.100	mm/min
σ_{cam}	DV, unmeasured	0.35	0–3.00	mm
r	state	150.00	20–260	mm
h	state	8.00	0.2–40	mm
T_h	state	1695.96	—	°C
T_c	state	1566.73	—	°C
T_m	state	1425.00	—	°C
m	state	100.0	0.5–100	kg
L	state	0	—	mm
V_a	state	1.000	0.05–2.50	mm/min
P_a	state	118.26	30–220	kW
m_p	state	0	—	kg

where r is the crystal radius in mm, c_h is the growth angle change per unit meniscus height in rad/mm and h_0 is the height at which the angle is zero. The argument of the tangent is clipped to ± 0.5 rad, which is never reached in normal operation and keeps the right-hand side finite if a student drives the meniscus to a wall. Both sides of (3.2) carry mm/min.

Equations (3.1) and (3.2) are the whole reason the application exists. There is no term in (3.2) that contains V_p alone, so $\partial \dot{r} / \partial V_p = 0$ whenever $h = h_0$. The application's own test suite verifies that

algebraically by stepping the pull rate 20 % above design at the equilibrium meniscus and confirming that $|\dot{r}| < 10^{-12}$ mm/min, then displacing the meniscus by 1 mm and confirming that $\dot{r}$ becomes negative. Two integrations, no direct path, and the loop pays for it in Section 3.4.

3.3.2 Growth Rate and the Structure Limit

Latent heat released at the interface leaves by conduction up the crystal and radiation from its surface, and a thinner crystal removes it over a shorter path, so the rate at which the melt can freeze rises as the radius falls. A hotter melt carries superheat into the interface and slows it. The growth rate retains those two terms only, as shown in Equation (3.3).

$$V_g = \left[V_0 \frac{r_0}{r} - K_T \left(T_m - T_{m0} \right) \right] \eta \tag{3.3}$$

where V_0 is the growth rate at the design radius r_0 and the design melt temperature T_{m0}, K_T is the growth rate lost per degree of melt superheat in mm/min per °C, and η is a hot-zone multiplier representing how well the heat shield removes heat up the crystal. The radius in the denominator is floored at 10 mm. The same heat-removal argument bounds the pull rate, as shown in Equation (3.4).

$$V_{\max} = V_{\max,0} \frac{r_0}{r} \eta \tag{3.4}$$

where $V_{\max,0}$ is the pull rate at which structure is lost at the design radius. Above $V_{\max}$ the interface cannot shed the latent heat, the growth front breaks down and the crystal ceases to be a single crystal. The application latches a loss flag at that instant and stops accruing prime crystal.

Equation (3.4) deserves its own sentence, because the modeling choice behind it is where this application differs from a textbook saturation exercise. $V_{\max}$ is a *process* constraint and not an actuator stop. It is not the limit the controller output is clamped to, it falls as the crystal fattens, no instrument in the application measures it, and the controller can and will drive straight through it. The pull drive has a separate and larger mechanical limit of 2.50 mm/min, and that is what the diameter controller output is clamped to. Declaring $V_{\max}$ as the controller output limit would make the constraint uncrossable and would remove the only failure the application is built around.

3.3.3 Melt Inventory, Length and Product

The mass solidified per minute follows from the interface area and the growth rate, as shown in Equation (3.5).

$$\dot{m}_s = \rho_{Si} \pi r^2 \max(0, V_g), \qquad \frac{\mathrm{d}m}{\mathrm{d}t} = -\dot{m}_s, \qquad \frac{\mathrm{d}L}{\mathrm{d}t} = V_p \tag{3.5}$$

where ρ_{Si} is the density of solid silicon in kg per mm^3, m is the melt remaining in kg and L is the crystal length in mm. At the design point (3.5) returns $\dot{m}_s = 0.1647$ kg/min, so the 88 kg between m_0 and m_{end} takes 534 min at exactly 1.000 mm/min, and makes 534 mm of boule. Prime crystal is integrated rather than counted, so the figure is exact at any step size, as shown in Equation (3.6).

$$\frac{\mathrm{d}m_p}{\mathrm{d}t} = \begin{cases} \dot{m}_s, & |D - D_{\mathrm{SP}}| \le \delta \ \wedge \ V_{lo} \le V_g \le V_{hi} \\ 0, & \text{otherwise or after structure is lost} \end{cases} \tag{3.6}$$

where δ is the diameter tolerance and $[V_{lo}, V_{hi}]$ is the growth rate band inside which the crystal freezes in neither an excess of vacancies nor an excess of self-interstitials.

3.3.4 The Thermal Chain

Heater power reaches the interface through three lumped nodes, as shown in Equation (3.7).

$$\begin{aligned} C_h \frac{\mathrm{d}T_h}{\mathrm{d}t} &= 60\,[P_a - U_{hc}(T_h - T_c) - U_{ha}(T_h - T_\infty)\lambda] \\ C_c \frac{\mathrm{d}T_c}{\mathrm{d}t} &= 60\,[U_{hc}(T_h - T_c) - U_{cm}(T_c - T_m) \\ &\qquad - U_{ca}(T_c - T_\infty)\lambda] \\ C_m \frac{\mathrm{d}T_m}{\mathrm{d}t} &= 60\,[U_{cm}(T_c - T_m) - U_{ma}(T_m - T_\infty)\lambda - q_{lat}] \end{aligned} \tag{3.7}$$

where C_h, C_c and C_m are heat capacities in kJ/K, the U terms are conductances in kW/K, T_∞ is the hot-zone reference temperature, λ is an insulation multiplier, and the factor 60 converts kW to kJ/min because the clock is in minutes. The latent heat drawn from the melt is $q_{lat} = L_f \dot{m}_s / 60$ in kW. Heater power is floored at zero before it is used.

Three of the coefficients in (3.7) are functions of the charge remaining, and they are not functions of it in the same way, as shown in Equation (3.8).

$$C_m = c_p \max(2, m), \;\; U_{cm} = U_{cm,0}\,\phi^{0.10}, \;\; U_{ma} = \text{constant} \tag{3.8}$$

where $\phi = \mathrm{clamp}(m/m_0, 0.02, 1.2)$ is the melt fraction and c_p is the melt heat capacity in kJ/(kg K). The heat capacity falls in proportion to the mass. The crucible-to-melt coupling follows the wetted wall, which falls much more slowly. The loss from the melt surface does not fall at all, because the surface area is set by the crucible diameter rather than by how much melt is in it. Section 3.4 measures what that combination does to the plant over a run.

3.3.5 Actuators, Instrument and Limits

Both actuators are first-order lags on their commands, as shown in Equation (3.9).

$$\tau_V \frac{\mathrm{d}V_a}{\mathrm{d}t} = V_{cmd} - V_a, \qquad \tau_P \frac{\mathrm{d}P_a}{\mathrm{d}t} = P_{cmd} - P_a \tag{3.9}$$

where V_a is the pull rate actually applied and P_a is the power actually delivered. The pull rate used everywhere in the physics is $V_p = \max(0, V_a)$. The bright-ring camera adds Gaussian noise to the true diameter and filters the result, as shown in Equation (3.10).

$$\tau_{cam} \frac{\mathrm{d}\hat{D}}{\mathrm{d}t} = 2r + \sigma_{cam}\, n(t) - \hat{D} \tag{3.10}$$

where $\hat{D}$ is the measured diameter in mm, σ_{cam} is the camera noise standard deviation and $n(t)$ is a standard normal deviate redrawn at every controller execution from the seeded generator. Advancing that filter uses the exact zero-order-hold recursion $\hat{D}_k = a\hat{D}_{k-1} + (1-a)(2r + \sigma_{cam} n)$ with $a = \exp(-\Delta t_c / \tau_{cam})$, which cannot go unstable however coarse the scan becomes.

After every step the radius is clipped to 20–260 mm, the meniscus to 0.2–40 mm and the melt to at least 0.5 kg. Three latches then run in order. Structure is lost when V_p has stayed above $V_{\max}$ continuously for t_{lose}, the crystal detaches if $h > h_{\max}$, and the crystal touches down into the melt if $h < h_{\min}$. Any of the three ends the useful run. The dwell matters more than it looks. Dislocation nucleation is not instantaneous, and without it one noisy camera sample differentiated through a 45 min derivative time scraps a boule, which is an artifact of the sample rate rather than of the interface. At $t_{lose} = 0.5$ min scenario 9 run with a 10 min derivative filter crosses $V_{\max}$ once, for three physics steps, and survives, while the same scenario as loaded stays across it for the whole dwell and loses the crystal. The dwell is accumulated as a running sum of physics steps, and the comparison allows for a sum that falls short of t_{lose} by a rounding error, so ten consecutive steps above the limit trip the latch. Separately, the batch is declared finished when $m \le m_{end}$, at which point the mechanical derivatives are frozen so that the model is never asked to place a meniscus in an empty crucible.

Table 3.2 Constants of the Puller Model. A Hidden Constant Has No Entry in the Parameter Panel and Has to Be Recovered From an Experiment.

Symbol	Meaning	Value	Unit	Visibility
h_0	meniscus at zero angle	8.0	mm	adjustable
c_h	angle per unit h	0.050	rad/mm	adjustable
$h_{\min}$	floods below	3.0	mm	adjustable
$h_{\max}$	detaches above	15.0	mm	adjustable
V_0	design growth rate	1.00	mm/min	adjustable
K_T	growth per degree	0.025	(mm/min)/K	adjustable
$V_{\max,0}$	structure limit at r_0	1.60	mm/min	adjustable
C_h	heater capacity	300	kJ/K	adjustable
C_c	crucible capacity	900	kJ/K	adjustable
U_{hc}	heater to crucible	0.80	kW/K	adjustable
$U_{cm,0}$	crucible to melt, full	0.60	kW/K	adjustable
U_{ma}	melt surface loss	0.356	kW/K	adjustable
U_{ca}	crucible loss	0.050	kW/K	adjustable
U_{ha}	heater loss	0.030	kW/K	adjustable
m_0	melt charge	100	kg	adjustable
m_{end}	melt left at end	12	kg	adjustable
t_{lose}	dwell above $V_{\max}$	0.50	min	adjustable
τ_V	pull drive lag	0.30	min	adjustable
τ_P	heater supply lag	0.50	min	adjustable
δ	diameter tolerance	1.5	mm	adjustable
V_{lo}	defect band, lower	0.72	mm/min	adjustable
V_{hi}	defect band, upper	1.32	mm/min	adjustable
—	value of prime crystal	900	\$/kg	adjustable
r_0	design radius	150	mm	hidden
T_{m0}	design melt temp.	1425	°C	hidden
T_{mp}	silicon melting point	1414	°C	hidden
c_p	melt heat capacity	1.00	kJ/kg/K	hidden
T_∞	hot-zone reference	1200	°C	hidden
L_f	latent heat of fusion	1800	kJ/kg	hidden
ρ_{Si}	solid silicon density	2.33e−6	kg/mm^3	hidden

3.3.6 Parameters, Initialization and Numerics

Table 3.2 lists every constant. Seven are hidden from the parameter panel and are the quantities a student has to infer from experiments rather than read.

Every scenario but the tail scenario starts from a condition in which nothing is moving. The radius is set to r_0, the meniscus to h_0, the melt to m_0 and the melt temperature to T_{m0}, and the two remaining temperatures are solved backward up the chain so that all three derivatives of (3.7) vanish, as shown in Equation (3.11).

$$\begin{aligned} q_m &= U_{ma}(T_{m0} - T_\infty)\lambda + L_f \dot{m}_s/60 \\ T_c(0) &= T_{m0} + q_m/U_{cm} \\ q_c &= q_m + U_{ca}(T_c(0) - T_\infty)\lambda \\ T_h(0) &= T_c(0) + q_c/U_{hc} \\ P_a(0) &= q_c + U_{ha}(T_h(0) - T_\infty)\lambda \end{aligned} \tag{3.11}$$

At the design point (3.11) returns $T_c = 1566.73\,^\circ\mathrm{C}$, $T_h = 1695.96\,^\circ\mathrm{C}$ and $P_a = 118.26\,\mathrm{kW}$, and the same expression supplies the design-point aid available in manual mode. The application's test suite confirms that the largest state derivative at that point is below 10^{-9} in its own units and that the diameter is exactly 300 mm. One scenario overrides the initial state, starting the run at $m = 30\,\mathrm{kg}$ and $L = 1100\,\mathrm{mm}$.

States advance under fixed-step fourth-order Runge-Kutta with a step of 0.05 min (3 s). DIC-301 and VIC-301 both run on a 0.1 min (6 s) period, and the historian keeps two samples for every simulated minute. The noise sequence comes from a generator seeded at 20260918, so scenario 9

produces the same camera record on any machine. Section 1.6.1 describes the engine, the seed and the speed setting, which leaves the trajectory untouched.

3.3.7 Assumptions

The model is detailed where the two-actuator structure lives and thin everywhere else, and each simplification fails somewhere.

One-dimensional quasi-steady heat transfer through three nodes. Real pullers are dominated by three-dimensional buoyant and Marangoni convection in the melt, which is why production machines apply a magnetic field to suppress it, and that convection supplies a large part of the real diameter disturbance [13]. The assumption fails wherever the melt flow pattern itself changes, which is exactly where a real diameter excursion starts.

The meniscus is linearized about the constant growth angle. Equation (3.2) is the small-departure form of the Laplace-Young solution rather than the solution itself [8], so the flooding and detachment heights in Table 3.2 are placed from experience rather than derived, and the model is not trustworthy for the large angle excursions of a shoulder.

Growth rate is two terms. There is no axial temperature gradient field in the model, so the defect specification is applied to V_g directly rather than to the ratio of growth rate to gradient that actually governs which point defects freeze in [5, 6]. The substitution is only valid while the gradient is treated as fixed, which a change of hot zone would break.

Loss of structure is instantaneous. In practice the interface breaks down over some minutes and an experienced operator sometimes sees it coming in the ring image, so the latch in this model is pessimistic about warning and optimistic about recovery.

The crucible lift is perfect unless disturbed. A real lift profile is itself a tuned trajectory that interacts with the thermal field, and here it is a constant error term in (3.1) with no dynamics. A lift error that appears instantly at 0.03 mm/min and never changes is not what a real lift mechanism does, but it isolates an unmeasured ramp into the one integrator that reaches the diameter. In the same way a hot-zone multiplier that changes in one instant is a step because a step is the only upset whose response a trend reads back cleanly, while a real argon disturbance arrives with a shape of its own.

No oxygen, carbon or dopant transport. Resistivity control, striations and axial segregation are all real control problems on the same machine and none of them is present, so the economics in the application are a flat value per kilogram rather than the graded price a real boule carries.

Repetition is exact, which no control system manages. Freezing the integration step and the scan makes one trajectory the only trajectory, where a real puller control system carries scan jitter, a coarser measurement resolution and an execution order that shifts.

3.4 Dynamic Analysis

3.4.1 The Diameter Plant Is a Double Integrator

Linearizing (3.1) and (3.2) about a steady operating point at which $h = h_0$ and $V_p = V_g$, and holding the melt temperature fixed so that V_g does not respond, gives $\Delta\dot{h} = \Delta V_p$ and $\Delta\dot{r} = -V_p c_h \Delta h$. Combining them gives the plant in Equation (3.12).

$$\frac{D(s)}{V_p(s)} = \frac{-2V_p c_h}{s^2 + \omega_n^2}, \qquad \omega_n = V_p \sqrt{\frac{c_h}{r_0}} \tag{3.12}$$

where the ω_n^2 term comes from the feedback of radius into growth rate through (3.3), and the leading factor of two converts radius to diameter. At the design point $\omega_n = 0.018\,26\,\text{rad/min}$, which is a natural period of 344 min, and the steady-state gain is $2r_0/V_p = 300\,\text{mm}$ of diameter per mm/min of pull rate. The negative sign means that pulling faster makes the crystal thinner, so a crystal growing fat is answered by pulling faster, and the diameter loop is direct acting in the convention of Section 1.6.2.

Two facts about (3.12) decide the tuning. Its phase begins at $-180°$ and stays there, so proportional and integral action alone leave no margin at any gain and the loop rings until something saturates.

Table 3.3 Derivative Time Against Phase Margin and Measured Diameter Excursion. Margins Are Computed From (3.12) at $V_p = 1\,\text{mm/min}$ With the Shipped Filter $\tau_F = 8\,\text{min}$. The Excursion Is the Peak-to-Peak Diameter Over $120 \le t \le 500\,\text{min}$ After a 4 mm Setpoint Step at $t = 60\,\text{min}$ in Scenario 3. Without Derivative Action the Meniscus Floods at $t = 158.4\,\text{min}$, So That Column Spans $60 \le t \le 158.4\,\text{min}$.

τ_D (min)	0	15	30	45	90
Phase margin (°)	−6.2	25.4	30.1	29.2	24.4
Crossover period (min)	82.0	59.4	44.3	36.3	25.6
Peak-to-peak D (mm)	11.28	2.52	1.98	2.29	3.50
Peak D (mm)	309.8	306.5	306.3	306.2	306.3
Meniscus floods	158.4 min	—	—	—	—

Derivative action is therefore the entire stabilizing mechanism rather than a refinement, as it is for the ball and beam of Volume 7, Chapter 5, and the amount required is set by where the derivative zero has to sit relative to crossover. Table 3.3 evaluates the loop $L(s) = K_c[1 + 1/(\tau_I s) + \tau_D s/(\tau_F s + 1)]\,D(s)/V_p(s)$ at the default $K_c = 0.055\,\text{mm/min}$ per mm, $\tau_I = 120\,\text{min}$ and $\tau_F = 8\,\text{min}$, beside the diameter excursion measured in the application on a 4 mm setpoint step.

The first column of Table 3.3 is a negative phase margin and the run does not survive it. Without derivative action the diameter overshoots the new setpoint to 309.8 mm at $t = 101.9\,\text{min}$, swings back to 298.5 mm at $t = 141.9\,\text{min}$, and the growing oscillation in the meniscus floods it at $t = 158.4\,\text{min}$, so the crystal is lost through the meniscus while the pull rate never comes closer than 0.214 mm/min to the structure limit. With $\tau_D = 45\,\text{min}$ the same step is answered with an excursion of 2.29 mm and a phase margin of 29.2°. Forty-five minutes of derivative time sounds enormous until the crossover period of 36.3 min is set beside it. The rule a double integrator enforces is that the derivative time has to be comparable to the closed-loop response time rather than a tenth of it.

🕮 **Pause.** Try Exercise 3.1 before reading on.

Those margins also price the derivative filter. The same three gains at $\tau_F = 3\,\text{min}$ give 50.9° at a crossover period of 28.4 min, so opening the filter to the shipped 8 min costs 21.7° of margin, on a plant that began with none.

3.4.2 An Exact Gain Schedule

Both the gain and the natural frequency of (3.12) are proportional to the pull rate, which is unusual and useful. Substituting $s = (V_p/V_0)\,\sigma$ into (3.12) gives Equation (3.13).

$$\left.\frac{D}{V_p}\right|_{V_p}(s) = \frac{V_0}{V_p}\cdot\frac{-2V_0 c_h}{\sigma^2 + \omega_{n,0}^2} = \frac{V_0}{V_p}\cdot\left.\frac{D}{V_p}\right|_{V_0}(\sigma) \tag{3.13}$$

where $\omega_{n,0} = V_0\sqrt{c_h/r_0}$ is the natural frequency at the design pull rate. The plant at pull rate V_p is therefore the design-point plant with its frequency axis stretched by V_p/V_0 and its magnitude reduced by the same factor. A controller that stretches its own frequency axis by exactly the same amount and raises its gain by the reciprocal of the magnitude change reproduces the loop shape identically, as shown in Equation (3.14).

$$K_c \leftarrow K_c\,\frac{V_p}{V_0}, \qquad (\tau_D,\ \tau_I,\ \tau_F) \leftarrow (\tau_D,\ \tau_I,\ \tau_F)\,\frac{V_0}{V_p} \tag{3.14}$$

Equation (3.14) is a schedule the plant equations hand the engineer rather than one fitted during commissioning, and it is exact rather than approximate. The distinction matters, because the general difficulty with gain scheduling is that a family of linear designs interpolated on a measured variable carries no guarantee away from the design points [18]. Here the scheduling variable is the manipulated

variable itself, it is measured exactly, and it is algebraically the reason the gain moves instead of a stand-in for the reason. Chapter 2 schedules its loop on a measured cause in the same way, the rod diameter. The application implements Equation (3.14) on the applied pull rate V_a, with the ratio V_a/V_0 clamped to 0.2–4 so that a camera excursion cannot run the gain away.

Pause. Try Exercise 3.3 before reading on.

Figure 3.2 evaluates the phase margin of the diameter loop against pull rate three ways. Applying (3.14) holds the margin at 29.19° for every pull rate from 0.5 to 2.0 mm/min and not merely close to it, which is what exact loop-shape preservation means. A fixed tuning slides from 38.1° at 0.5 mm/min to 21.3° at 2.0 mm/min. The third curve is the reason the direction has to be derived rather than guessed. Scaling K_c by V_0/V_p instead, together with τ_D and with τ_I and τ_F left fixed, which is the intuitive choice because a faster plant seems to want a softer controller, gives 24.5° at 0.5 mm/min, worse than no schedule at all over the whole lower half of the drive range. The plant is faster at high pull rate and also weaker, and the gain has to follow the weakness.

A measured step confirms the shape. Settling the full charge for 250 min at mid-ranging targets of 1.00, 1.40 and 0.70 mm/min with the camera noise off and then stepping the diameter setpoint 3 mm gives a highest diameter over the following 120 min of 303.71 mm unscheduled and 303.70 mm scheduled at $V_p = 1.044\,\mathrm{mm/min}$, because the schedule is within 5 % of unity there. At $V_p = 1.343\,\mathrm{mm/min}$ the fixed tuning already sits 2.53 mm high before the step and peaks at 305.44 mm, against 1.45 mm high and a peak of 304.43 mm scheduled. At $V_p = 0.822\,\mathrm{mm/min}$ the scheduled loop is the slower of the two, reaching 301.81 mm in 120 min against 302.55 mm, and that is correct rather than a failure, because (3.13) stretches the frequency axis with the pull rate and what the schedule preserves is damping, not speed.

The flatness in Figure 3.2 is exact only because every time constant in that loop scales with the pull rate, and two of the real ones do not. The bright-ring camera filter of 0.50 min in (3.10) and the pull drive lag τ_V of 0.30 min in (3.9) both sit inside the loop and neither is a controller setting, so (3.14) cannot reach them. Adding the camera lag alone to the loop transfer function takes the unscheduled margin from 34.9 to 14.1° over the same 0.5 to 2.0 mm/min sweep and the scheduled margin from 26.7 to 19.6°. Adding the drive lag as well gives 33.1 to 9.7° unscheduled against 25.2 to 13.9° scheduled, so the honest margin of the complete loop at the design point is 21.4° rather than the 29.2° of the meniscus loop alone. Nothing is flat once the fixed lags are in, and the schedule's worth has to be restated because of it. What it buys is margin where the loop is weakest: at 2.0 mm/min it takes the complete loop from 9.7 to 13.9°, which is the difference between a loop that rings on every disturbance and one that does not. That is a more defensible claim than the flatness and it is the one that matters in operation, because the fast end is where a mid-ranged pull rate spends its time after a disturbance.

The boundary is worth stating in general terms. A gain schedule preserves loop shape only over the dynamics whose time constants it sets, so the residue is exactly what it cannot reach, and no amount of scheduling converts it into authority. Scenario 8 shows the second half of that plainly. At a mid-ranging target of 1.35 mm/min the diameter walks from 300.0 to 303.5 mm and leaves the 1.5 mm band at $t = 50.3$ min unscheduled and 52.3 min scheduled, with a worst true-diameter deviation of 3.46 mm unscheduled and 2.73 mm scheduled. The schedule narrows the excursion by a fifth and does not close it, because the loop has been given back its shape and not its authority.

3.4.3 The Heater Path and Why It Is the Outer Loop

The heater reaches the diameter only by changing the growth rate, which the diameter loop then has to match with the pull rate. Linearizing the thermal chain (3.7) together with the latent-heat feedback at the design point gives eigenvalues at -0.5342, -0.2161 and -0.0118/min, which are time constants of 1.87, 4.63 and 84.8 min. The slow mode dominates because the three capacities total 1300 kJ/K while the net conductance to the hot zone is only 0.3125 kW/K once the latent-heat feedback of 0.1235 kW/K is credited against the 0.436 kW/K of ambient loss. That lumped estimate is 69 min,

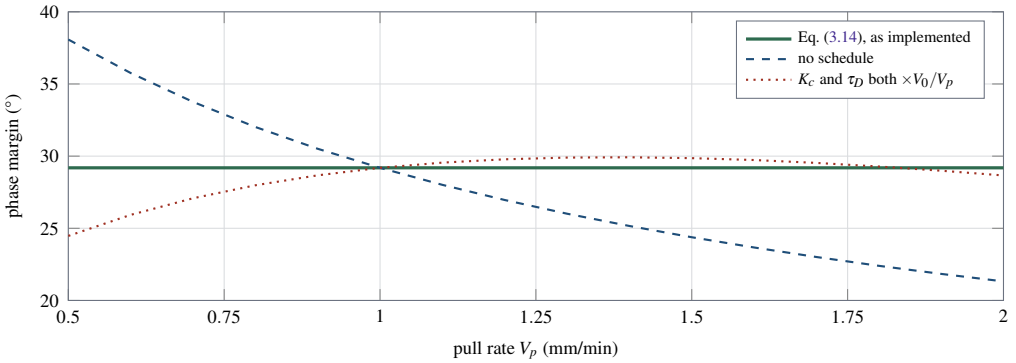


Figure 3.2 Phase margin of the diameter loop against pull rate for three scheduling choices, evaluated on Equation (3.12). All three curves start from the default tuning $K_c = 0.055\,\text{mm/min}$ per mm, $\tau_I = 120\,\text{min}$, $\tau_D = 45\,\text{min}$ and $\tau_F = 8\,\text{min}$ at $V_p = V_0$. The schedule of Equation (3.14) is the thick solid line, no schedule the dashed line and the reversed scaling the dotted line. Only Equation (3.14) holds the margin, because only it stretches the controller frequency axis and raises the gain together. Scaling the gain and the derivative time the other way, with τ_I and τ_F fixed, is worse than no schedule below the design pull rate.

short of the 84.8 min the full linearization returns but the right mechanism. On the linearized chain at the full charge a 5 kW heater step raises the melt by 14.09 °C, reaching 63 % of that change in 91.4 min and 95 % in 260.5 min, which is the slow mode with the two fast ones in series ahead of it. The application cannot show the whole of that response with nothing holding the diameter. Stepping the heater 5 kW in manual mode from the 118 kW at which scenario 1 opens warms the melt 9.39 °C above an otherwise identical unstepped run by $t = 88.3\,\text{min}$, when the slower growth has stretched the meniscus past $h_{\max}$ and the run stops. The linearized chain gives 8.72 °C at the same instant, and the difference is the latent heat no longer drawn by a crystal that has narrowed to 277 mm.

The static path from power to pull rate follows from (3.3). A change of 1 kW moves the melt by 2.819 °C, which moves the growth rate by $-K_T \times 2.819 = -0.0705\,\text{mm/min}$, and at steady state the diameter loop must set $V_p = V_g$. With the mid-ranging gain at 8 kW per mm/min the loop gain is $8 \times 0.0705 = 0.564$, so with $\tau_I = 120\,\text{min}$ the mid-ranging loop returns the pull rate with a time constant near $120/0.564 = 213\,\text{min}$. That is 5.9 times the diameter loop crossover period of 36.3 min, which is the separation the structure needs. The mid-ranging loop acts *through* the diameter loop, so if it were the faster of the two the loops would chase the same meniscus and the crystal would carry a slow waist. Cascade tuning states the rule as inner fast and outer slow. Mid-ranging is the same rule with the roles of speed and authority exchanged, and the outer loop is the one with the authority.

3.4.4 A Batch Has No Steady State

Nothing in the application is stationary, because the charge drains continuously and a completed run takes between 440 and 520 min depending on how fast the diameter loop is pulling. The batch crystallizer of Volume 8, Chapter 4 likewise follows a trajectory rather than a steady state. Using the melt node alone, the time constant is $C_m/[60(U_{cm} + U_{ma})]$ and the static gain from the crucible is $U_{cm}/(U_{cm} + U_{ma})$. Evaluating both from (3.8) gives 1.743 min and 0.628 at the full 100 kg charge against 0.293 min and 0.582 at 15 kg. The melt therefore answers 5.9 times faster at the tail than at the shoulder while its gain changes by only 7.2 %, which is an unusually clean separation of speed from gain and the reason a single mid-ranging tuning is wrong for part of every run.

The consequence for an operator is sharper than the arithmetic suggests. Scenario 1 opens in manual mode with the pull rate at the design 1.000 mm/min and the heater at 118 kW, within 0.26 kW of the design power, disturbs nothing, and leaves them there. Over the following four hours the

melt cools from 1425.0 to 1418.0 °C because the crucible-to-melt coupling is shrinking, the crystal therefore solidifies slightly faster than it is being pulled, the meniscus shortens from 8.00 to 3.39 mm, the diameter grows to 354.7 mm, and the crystal touches down into the melt at $t = 256.3$ min. Setting the heater to the 118.3 kW that the design-point aid computes only postpones the touch-down to $t = 312.0$ min. A perfectly executed recipe scraps the run.

3.5 Control Structure

3.5.1 The Shared PID Algorithm

Both controllers are the velocity-form PID of Section 1.6.2, with the derivative on the filtered measurement and back-calculation anti-windup on the integral alone. Equation (3.15) writes it in this chapter's symbols: γ is the action a of Equation (1.12), Σ_k the running sum S_k and u_k^* the output before limiting v_k.

$$
\begin{aligned}
e_k &= \gamma(\mathrm{SP}_k - \mathrm{PV}_k) \\
P_k &= K_c e_k, \qquad D_k = -K_c \tau_D \gamma \frac{\mathrm{PV}_k^f - \mathrm{PV}_{k-1}^f}{\Delta t_c} \\
\Sigma_k &= \Sigma_{k-1} + (P_k - P_{k-1}) + (D_k - D_{k-1}) \\
u_k^* &= u_{bias} + \Sigma_k + I_{k-1} \\
u_k &= \min(u_{\max}, \max(u_{\min}, u_k^*)) \\
I_k &= I_{k-1} + \left[\frac{K_c}{\tau_I} e_k + \frac{u_k - u_k^*}{T_t} \right] \Delta t_c
\end{aligned}
\tag{3.15}
$$

where $\gamma = \pm 1$ is the action ($+1$ reverse acting and -1 direct acting), PV^f is the measurement after a first-order filter of constant τ_F advanced by the same zero-order-hold recursion as the camera, Δt_c is the controller period and T_t is the anti-windup tracking constant, which defaults to τ_I. With the schedule of Equation (3.14) on, T_t follows the scheduled τ_I. At $T_t = 120$ min and $\Delta t_c = 0.1$ min each scan would remove $1/1200$ of the clipped excess, but neither loop reaches its output limits in any run reported here, so the tracking time affects none of the results.

Transfer between manual and automatic is bumpless as Section 1.6.2 describes, and the derivative estimate is kept warm in manual so that it is correct the instant automatic is selected. Neither loop carries a rate limit.

3.5.2 The Two Loops as Configured

Diameter controller DIC-301 is direct acting, $\gamma = -1$, because opening the pull rate makes the crystal thinner, in the convention of Section 1.6.2. Its setpoint is D_{SP}, its measurement is $\hat{D}$ from (3.10), and its output is clamped to the pull drive limits of 0.05 to 2.50 mm/min. The defaults are $K_c = 0.055$ mm/min per mm, $\tau_I = 120$ min, $\tau_D = 45$ min and $\tau_F = 8$ min, with the camera filter $\tau_{cam} = 0.50$ min ahead of it.

Mid-ranging controller VIC-301 is also direct acting, $\gamma = -1$, because more heater power slows the growth and the diameter loop answers by pulling more slowly, so a pull rate above its target calls for more heat. Its setpoint is V_{SP}, the pull rate the heater loop aims to leave the drive at. Its measurement is not the diameter and not the controller output either, but the pull rate actually applied, V_p, which is the state of (3.9). Its output is clamped to 30 to 220 kW, and its defaults are $K_c = 8$ kW per mm/min and $\tau_I = 120$ min with no derivative action. The mid-ranging loop can be switched out, in which case the heater holds its last command and cannot be moved by hand. Manual mode hands both the pull rate and the heater to the operator.

🕮 **Pause.** Try Exercise 3.2 before reading on.

The optional schedule applies Equation (3.14) on every controller execution, multiplying K_c by V_a/V_0 and dividing τ_I, τ_D and τ_F by the same ratio, with V_a floored at 0.05 mm/min and the ratio clamped to 0.2–4. Figure 3.2 reports what it does to the margin and Section 3.6 what it does on the

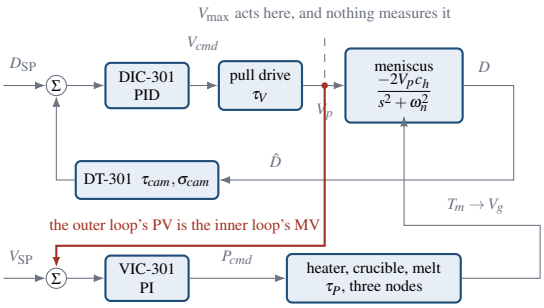


Figure 3.3 The mid-ranging structure as implemented. DIC-301 holds the diameter with the pull rate through a double integrator. VIC-301 never sees the diameter, takes the pull rate as its measurement, and walks the heater until the pull rate returns to V_{SP}. The unmeasured constraint V_{max} acts on the same signal VIC-301 regulates.

scenarios. Figure 3.3 shows the two loops and the single point at which they meet, which is the pull rate.

3.5.3 Model Predictive Control

An optional fourth structure, offered last in the Control list and loaded in charge only by scenario 11, replaces both loops in automatic mode with one model predictive controller (MPC) that moves the pull rate and the heater together. Every $T_s = 0.5$ min from $t = 0$ it solves one quadratic program over $N = 120$ steps (60 min) in move blocks of $\{1,1,1,1,2,2,3,4,6,9,15,75\}$ steps, then holds its first move until the following decision [19, 20], in the setup of Section 1.6.3. A 5 kW heater step moves the diameter only 0.88 mm in 30 min but 8.80 mm in 60 min.

The controller's model is not the puller of Section 3.3. It is the linearization of Equations (3.1) to (3.10) about the design point with the declared constants of Table 3.2 and the design disturbances ($\eta = \lambda = 1$, $\varepsilon_L = 0$), never the values of the run, so a constant a learner changes becomes model error hidden from the controller. Its nine states, written as departures from the design point with a tilde, are V_a, P_a, the three temperatures, h, r, the filtered diameter and one disturbance d, and its inputs are V_{cmd} and P_{cmd}. The lags are (3.9), the camera is (3.10) at its declared τ_{cam} without noise, the thermal chain is (3.7) frozen at the full charge ($C_m = c_p m_0$, $U_{cm} = U_{cm,0}$), and the remaining rows are Equation (3.16).

$$\begin{aligned} \frac{\mathrm{d}\tilde{h}}{\mathrm{d}t} &= \tilde{V}_a - \tilde{V}_g + d, \qquad \frac{\mathrm{d}\tilde{r}}{\mathrm{d}t} = -\bar{V} c_h \tilde{h}, \qquad \frac{\mathrm{d}d}{\mathrm{d}t} = 0 \\ \tilde{V}_g &= -\frac{V_0 r_0}{\bar{r}^2}\,\tilde{r} - K_T\,\tilde{T}_m \\ \tilde{q}_{lat} &= \frac{L_f \rho_{Si} \pi}{60}\left(2 r_0 V_0\,\tilde{r} + r_0^2\,\tilde{V}_g\right) \end{aligned} \tag{3.16}$$

where $\tilde{q}_{lat}$ enters the melt row of (3.7), and the radius path's operating point $\bar{V}$, $\bar{r}$ is reset at every decision to the measured V_a and $\hat{D}/2$, because the pull rate moves the gain and natural frequency of (3.12). The disturbance d carries, in growth-rate units, whatever the declared model omits, such as a lift error, a hot-zone change ($-V_g\,\Delta\eta$) or a draining melt's cooling ($K_T\,\Delta T_m$). An exact zero-order hold discretizes the model.

A central-difference Jacobian of the plant's derivative function at the design point agrees with every entry of the model to 9.8×10^{-11}, and its meniscus and radius pair, damped by the latent heat, sits at $-0.00149 \pm 0.01609\,i$ per minute. Stepped in manual mode against an unstepped run, noise off, from 30 to 120 min, the model's diameter is within 1.0 to 3.8 % of the plant's for pull-rate steps

of +0.01 and −0.05 mm/min and 3.2 to 6.9 % short for heater steps of ±2 and +10 kW, because the plant's melt drains.

Offset is removed by estimating d [21]. A steady-state Kalman filter at the 0.1 min controller period reads the two signals the shipped loops read, $\hat{D}$ at a variance of 0.012 mm^2 and V_a at 10^{-6}, with process noise of 10^{-5} on h, r and d only. On the model a step in d is 63 % estimated in 10 min, the slowest error mode is 84.7 min, and the filter runs under every structure. The estimate of d is the controller's only integral action. Each decision minimizes Equation (3.17) subject to Equation (3.18).

$$\begin{aligned} J = {} & \sum_{k=1}^{N} \left(2r_k - D_{\mathrm{SP}}\right)^2 + r_V \sum_{k=0}^{N-1} \left(V_{cmd,k} - V_{\mathrm{SP}}\right)^2 \\ & + \sum_{j=1}^{12} \left(s_V\, \Delta V_{cmd,j}^2 + s_P\, \Delta P_{cmd,j}^2\right) + \rho_D\, \varepsilon_D^2 + \rho_h\, \varepsilon_h^2 \end{aligned} \tag{3.17}$$

$$\begin{aligned} & 0.05 \le V_{cmd,j} \le \min\left(2.50,\ \hat{V}_{\max} - \delta_V\right) \\ & 30 \le P_{cmd,j} \le 220, \qquad \left|2r_k - D_{\mathrm{SP}}\right| \le \delta + \varepsilon_D \\ & h_{\min} + m_h - \varepsilon_h \le h_k \le h_{\max} - m_h + \varepsilon_h \end{aligned} \tag{3.18}$$

Here k counts prediction steps and j move blocks, the predicted true diameter $2r_k$ is weighted at 1 per mm^2, $r_V = 25$ per (mm/min)^2 prices the pull command's distance from V_{SP}, the target VIC-301 uses, and $s_V = 200$ per (mm/min)^2 and $s_P = 5$ per kW^2 price the change between blocks, the first from the commands applied now, with no terminal weight. The term in r_V is mid-ranging written as a price, and the heater has no reference of its own. Both input limits are hard and equal the loops' clamps, except that the pull rate is also held $\delta_V = 0.02\,\mathrm{mm/min}$ under $\hat{V}_{\max} = V_{\max,0}\, r_0/(\hat{D}/2)$, the controller's own estimate of (3.4) at the design hot zone, which the shipped loops are never given. The diameter band uses the tolerance δ at $\rho_D = 100$, and the meniscus band keeps the predicted meniscus $m_h = 1.5\,\mathrm{mm}$ inside the declared flooding and detachment heights at $\rho_h = 10^4$, both per mm^2. Each band has one nonnegative slack (ε_D, ε_h) for the whole horizon and is judged at every second step, so zero slack predicts a respected band without guaranteeing it at every step. The 24 block moves and two slacks are found by a dual active-set method, whose unconstrained first moves match independently formed normal equations to 1.2×10^{-11}.

A plan is rejected unless the solver reports it optimal with every move, prediction and cost finite and within bounds. On a rejection the commands hold for that scan, and DIC-301 and VIC-301, which track the applied commands, take over from the next scan until a valid plan returns. A rejection forced at 70.75 min of scenario 11 with the noise off moves neither command, and the loops then move at most 0.0076 mm/min and 0.050 kW per scan. The controller sees only $\hat{D}$, V_a, its commands, the setpoints, its settings and the declared constants, and a decision takes 0.40 to 0.47 ms on average in Node.js.

3.6 Results

Table 3.4 lists twelve scenarios in the order a laboratory session would take them. Each one pins the opening mode, the control architecture, the tuning constants, the four disturbance sliders and whatever timed events belong to it. Every figure reported below was read from the application's own scorecard using the physics, controller and logging periods of Section 3.3, and each run is carried to the end of the batch at $m = m_{end}$ or to the loss of structure, whichever arrives first, so an accumulated quantity such as IAE always carries its window with it. Pull-rate margin is the minimum of $V_{\max} - V_p$ over the run, evaluated at every 0.05 min physics step rather than at the two samples per minute the trend draws, because the quantity of interest is the closest the crystal actually came to losing its structure and not the closest a trend happens to show. The two differ by 0.018 to 0.025 mm/min on the runs of Table 3.5, always in the direction that flatters the loop. Every run starts from the scenario

Table 3.4 The Twelve Scenarios and What Each One Puts in Front of a Student

#	Scenario	What the student meets
1	The puller by hand	no steady state, meniscus flooding
2	Steady body growth	a plant that moves with nothing disturbed
3	4 mm setpoint step	derivative action on a double integrator
4	Argon upset, hot zone 1.04×	disturbance paid for in actuator position
5	Leaky hot zone plus argon upset	the structure limit the loop cannot see
6	Same run with mid-ranging	authority returned to the fast actuator
7	Crucible lift drifts –0.03 mm/min	an unmeasured ramp into the meniscus
8	Tail of the batch, $V_{SP} = 1.35\,\text{mm/min}$	a loop whose own output moves its plant
9	Camera noise 1.05 mm, $\tau_F = 4\,\text{min}$	derivative action on an optical measurement
10	Mid-range target 1.25 mm/min	an actuator that is also a specification
11	Two loops against one plan	both actuators in one predictive plan
12	Graded challenge, two upsets	the whole structure under a target

Table 3.5 Measured Outcome of Four Disturbance Sequences With the Diameter Loop Alone and With Mid-Ranging Added. Each Pair Is the Same Run With the Same Seed. Margin Is the Minimum of $V_{max} - V_p$ Over the Run at the 0.05 min Physics Step. Max $|\Delta D|$ Is on the Measured Diameter, Which Is What Both Loops and the Scorecard See, While the Prime Mass of (3.6) Is Scored on the True Diameter. The Last Pair Runs Scenario 9 With the Derivative Filter at 20 min.

Sc.	Heater	Margin (mm/min)	Max $\lvert\Delta D\rvert$ (mm)	Prime (kg)	Grown (kg)	Outcome
4	fixed	0.215	2.328	76.43	88.01	complete
4	mid-ranged	0.381	1.251	88.02	88.02	complete
5	fixed	0.221	2.349	73.46	88.00	complete
6	mid-ranged	0.358	1.375	88.00	88.00	complete
7	fixed	0.152	2.390	80.98	88.01	complete
7	mid-ranged	0.351	1.128	88.01	88.01	complete
9	fixed	−0.105	3.038	85.18	87.47	lost 483.0 min
9	mid-ranged	0.145	2.191	88.02	88.02	complete

as loaded, with any changed setting made at $t = 0$ before the run begins, which gives the same run as the setting carried through Reset.

3.6.1 The Loop Works and the Run Is Lost

Scenario 4 applies a hot-zone change of 1.04× at $t = 60\,\text{min}$, which makes the crystal solidify faster than it is being pulled. The meniscus shortens, the growth angle rises, the crystal fattens, and the diameter loop answers by pulling faster until growth and pull match again. The diameter is held 1.0 to 1.5 mm over 300 mm from $t = 130\,\text{min}$ to 390 min, inside its band, against 0.4 to 1.1 mm over in the same run without the upset, and everything looks resolved. It is not resolved. The pull rate is permanently higher, the margin to V_{max} is permanently smaller, and nothing in the interface reports that as a problem.

Table 3.5 runs four disturbance sequences with and without the mid-ranging loop. Every pair shares its disturbances, its tuning and its seed, and differs only in whether the heater is driven by the pull rate. Figure 3.4 draws the margin and the prime crystal of each pair.

Scenario 5 is the case the application is built around. The hot zone leaks 12 % more heat than the recipe assumed and an argon upset arrives on top of it at $t = 120\,\text{min}$. The diameter loop holds the diameter with a maximum deviation of 2.349 mm and never saturates, the pull rate ends the run at 1.378 mm/min against a structure limit of 1.651 mm/min, and the worst margin over the run is 0.221 mm/min, reached at $t = 447.3\,\text{min}$. Structure is never lost. The crystal survives, and only 73.46 kg of the 88.00 kg grown is prime, a yield of 83.5 %. Scenario 6 is the identical run with the

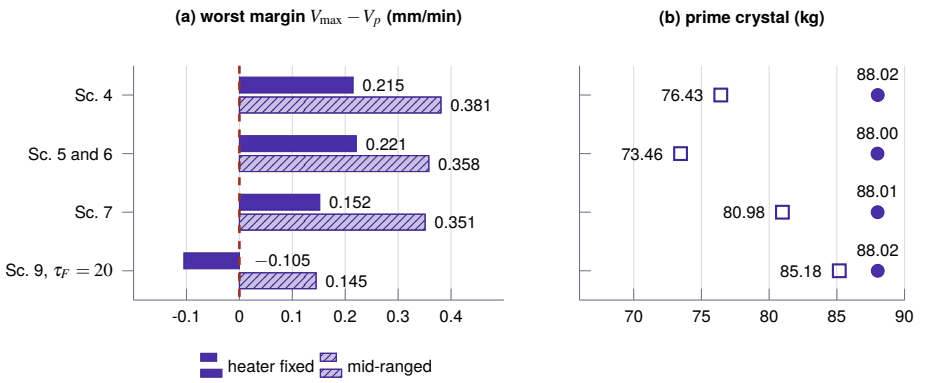


Figure 3.4 The four pairs of Table 3.5, each the same run with the heater fixed and with it mid-ranged. (a) The worst pull-rate margin over the run: solid bars for the fixed heater, hatched bars for the mid-ranged one, and the dashed line at zero, where the pull rate meets the structure limit. (b) The prime crystal: open squares for the fixed heater, filled circles for the mid-ranged one. Notice that mid-ranging widens the margin in every pair and keeps the whole of each crystal prime, while the fixed heater gives up prime crystal even in the runs that keep their structure.

heater driven by the pull rate. The margin never falls below 0.358 mm/min, the yield is 100 %, the maximum diameter deviation falls from 2.349 to 1.375 mm, and all 88.00 kg is prime, which at the flat valuation the application carries is \$79,200 against \$66,116. The diameter traces of scenarios 5 and 6 look alike, and their outcomes differ by 14.54 kg of product, so the actuator and not the controlled variable is the trace that tells them apart. The heater ends that run at 138.2 kW having started at 133.8 kW.

> **TRY IT IN THE APP** **5 · The Limit the Pull Rate Walks Into**
>
> Load the scenario and press Run. Watch the pull-rate strip: the pull rate climbs toward the red dashed structure limit while the diameter trace changes little. The closest approach, 0.221 mm/min, comes at 447.3 min, and only 73.46 kg of the 88.00 kg grown is prime. Reset, set Control to Diameter + mid-ranging and run again: the margin never falls below 0.358 mm/min, and all 88.00 kg is prime.

Scenario 7 makes the same point through a mechanism nobody is watching. A crucible lift that under-compensates by 0.03 mm/min is three percent of the pull rate, it enters the pure integrator of (3.1) directly, and no instrument in the application sees it. The diameter loop holds the diameter by offsetting the pull rate further and further, ending the run at 1.315 mm/min with a worst margin of 0.152 mm/min and 80.98 kg of 88.01 kg prime. With mid-ranging on, the heater absorbs the drift instead, moving from 118.26 to 121.59 kW over the run across a logged span of 117.56 to 122.09 kW, while the pull rate ends at 1.133 mm/min, the worst margin more than doubles to 0.351 mm/min, and the whole 88.01 kg is prime. A mid-ranged actuator that ends a batch somewhere it did not start is a diagnostic rather than a fault, because something upstream moved and reading the slow loop is often the only record of it.

The fourth pair is the one that loses a crystal. Scenario 9 degrades the camera to $\sigma_{cam} = 1.05$ mm, and with the derivative filter opened to 20 min so that the noise is not itself the failure, the diameter loop alone still walks the pull rate across V_{max} and holds it there past the dwell at $t = 483.0$ min, with 85.18 kg of 87.47 kg prime and everything after that instant polycrystalline. The same run mid-ranged finishes the batch at $t = 506.1$ min with a worst margin of 0.145 mm/min and 88.02 kg prime. Nothing about the diameter loop failed in the first of those runs. It held the diameter to within 3.038 mm using the only actuator it had, until that actuator ran out. The loss itself belongs to the shipped noise

Table 3.6 Noise Injected Into the Pull Rate by the Derivative Term, as the RMS Difference Between Paired Runs With and Without Camera Noise Over $60 \le t \le 240$ min in Scenario 6 With the Timed Upset Removed. 8 min Is the Default Filter, Which Every Scenario but Scenario 9 Ships With.

σ_{cam} (mm)	τ_F (min)	τ_D (min)	RMS ΔV_{cmd} (mm/min)	RMS ΔD (mm)
0.35	8	45	0.0429	0.0475
1.05	8	45	0.1288	0.1419
1.05	3	45	0.2886	0.1194
1.05	20	45	0.0650	0.1764
1.05	8	0	0.0275	0.3738

sequence. Over eight other seeds the diameter loop alone keeps its structure, with worst margins of -0.0004 to 0.174 mm/min and 83.9 to 85.8 kg prime, while the mid-ranged run keeps 88.0 kg prime and a worst margin of at least 0.186 mm/min on every seed, so the prime difference is robust and the loss is not.

3.6.2 Derivative Action Against an Optical Measurement

The price of the 45 min derivative time is paid on the camera. Table 3.6 measures it by running each case twice with identical seeds, once with the camera noise on and once with it off, and reporting the root-mean-square difference between the two trajectories over $60 \le t \le 240$ min. The differencing removes the drift of the batch and leaves only what the noise injected.

At the default filter of 8 min the camera noise carried into the pull-rate command rises from 0.0275 mm/min RMS without derivative action to 0.1288 mm/min with it, and to 0.2886 mm/min at $\tau_F = 3$ min. Recording the proportional and derivative terms of DIC-301 separately in the same paired runs isolates the derivative's own share, which is 5.4 times the proportional term's against the ratio $\tau_D/\tau_F = 5.6$ that the filtered derivative predicts, 13.7 times at $\tau_F = 3$ min against 15.0, and 2.2 times at 20 min against 2.25. The agreement is the point. Derivative noise gain on this loop is arithmetic, not a tuning surprise. The crystal itself barely registers any of it, because the meniscus integrates the command twice and the RMS diameter perturbation at the default filter is only 0.1419 mm, but the pull drive registers it and so does anybody reading the trend. The last row is not an option. With $\tau_D = 0$ the command noise falls to 0.0275 mm/min, but the loop has no phase margin and its own oscillation grows through the batch. The two runs of that pair drift apart as their oscillations grow, which is why its diameter difference is the largest in the table, and run to the end of the batch the noisy case leaves the diameter band at $t = 351.65$ min with a worst measured deviation of 3.99 mm.

Lengthening the filter is not free either. Going from 8 to 20 min halves the injected command noise to 0.0650 mm/min and adds phase lag exactly where (3.12) has none to spare, so the RMS diameter perturbation rises from 0.1419 to 0.1764 mm, and Section 3.4.1 prices the 3 to 8 min half of that move at 21.7° of phase margin. Scenario 9 ships with the schedule enabled, the camera degraded to $\sigma_{cam} = 1.05$ mm and the derivative filter shortened to 4 min, a noise gain of $\tau_D/\tau_F = 11.25$ that keeps the crystal on every one of 1001 noise seeds with the clean 0.35 mm camera. As configured the pull rate spikes to 1.79 mm/min, crosses the structure limit and stays there until the crystal is lost at $t = 113.3$ min, with 18.96 kg of prime crystal grown before that instant, and 1000 of 1001 seeds lose it. Raising the derivative filter to 12 min, the setting the scenario asks for, rescues the shipped run, which completes at $t = 506.1$ min with 88.01 kg prime but comes within 0.044 mm/min of the limit, and the rescue holds on 986 of 1001 seeds. That change costs 23.0° of phase margin on the loop of Table 3.3, from 44.3 to 21.4°. At the 8 min filter the application uses elsewhere the shipped seed also loses the crystal, at $t = 333.45$ min, but only 43 of 201 seeds do, and 20 min completes the shipped run with a worst margin of 0.145 mm/min. That trade is the permanent cost of stabilizing a double integrator with derivative action, and it is why a real puller spends its engineering on the camera rather than on the controller.

TRY IT IN THE APP 9 · A Noisy Bright Ring Against a Long Derivative Time

Run it as loaded. The pull rate spikes to 1.79 mm/min, crosses the structure limit and stays there, and the crystal is lost at 113.3 min with 18.96 kg of prime crystal. Reset, switch the panel to Advanced, set Derivative filter to 12 min and run again: the batch completes at 506.1 min with 88.01 kg prime, though the pull rate comes within 0.044 mm/min of the limit.

3.6.3 An Actuator That Is Also a Specification

The ratio of growth rate to the axial temperature gradient at the interface decides whether the crystal freezes in an excess of vacancies or an excess of self-interstitials, and only a narrow band of that ratio produces material free enough of both to make a device wafer [5]. The prediction was confirmed experimentally by varying the gradient and locating the defect boundaries against it [6]. The application applies the resulting specification directly to the growth rate through the band $[V_{lo}, V_{hi}] = [0.72, 1.32]\,\mathrm{mm/min}$ in (3.6).

Scenario 10 asks the mid-ranging loop to leave the pull rate at 1.25 mm/min instead of 1.00 mm/min for throughput. The diameter is held to a maximum deviation of 1.770 mm, the batch completes 65.9 min earlier at $t = 440.0\,\mathrm{min}$, and prime crystal falls from 88.01 to 74.51 kg. The growth rate leaves the band at $t = 385.2\,\mathrm{min}$, and 12.2 kg of the loss is grown above 1.32 mm/min from then on, while the other 1.2 kg is lost to a diameter excursion between 141 and 167 min. Throughput bought 65.9 min and sold 13.49 kg. The same run also drops its worst margin from 0.329 to 0.008 mm/min. With the diameter loop alone the heater stays at its design power and the target no longer acts, and the margin stays at or above 0.147 mm/min with 87.02 kg of 88.01 kg prime, so it is the throughput target and not the mid-ranging loop that spends the margin. An actuator that is simultaneously the manipulated variable of the primary loop and a controlled variable with a two-sided specification is an awkward object, and mid-ranging is what makes it tractable, because the diameter loop is free to move the pull rate in the short term while the heater sets where its average sits.

3.6.4 Model Predictive Control Against the Two Loops

Scenario 11 is the leaky hot zone of scenario 6 with the diameter setpoint stepped from 300 to 304 mm at 60 min and an argon upset at 240 min, loaded with the controller of Section 3.5.3 in charge. Table 3.7 runs it, scenario 3 and the ceiling case of scenarios 5 and 6 under the shipped loops, the same loops retuned once on scenario 11 and the controller as scenario 11 ships it. One rule chose the controller's weights from 300 candidates and the retuned loops by pattern search, the least true-diameter IAE on scenario 11 with no more travel on either actuator and no less worst margin than the shipped loops. For the loops it gives the schedule on, $K_c = 0.055$, $\tau_I = 20\,\mathrm{min}$ (its lower limit), $\tau_D = 30\,\mathrm{min}$ and $\tau_F = 8\,\mathrm{min}$ on DIC-301 and $K_c = 10\,\mathrm{kW}$ per mm/min, $\tau_I = 55\,\mathrm{min}$ on VIC-301. Nothing is retuned per case, each triple shares its seed and events, neither structure receives a preview, and each IAE has its own batch window. The loop comparison selects Diameter + mid-ranging explicitly in scenario 3, a case that ships with Diameter only, so "shipped loops" here names the gains, not the initial architecture of scenario 3.

Against the shipped loops the controller lowers the diameter IAE from 378.5 to 81.0 mm min on scenario 3, from 252.0 to 71.1 on the ceiling case and from 462.0 to 119.1 on scenario 11. Most of that gap is the shipped tuning. The retuned loops reach 149.6, 74.3 and 176.0 mm min, closing 77, 98 and 83 % of it, mainly through a 20 min integral time that keeps pace with the draining melt, and keep more margin than the controller on all three cases (0.338, 0.451 and 0.325 mm/min against 0.325, 0.403 and 0.313). No run in Table 3.7 comes within 0.08 mm/min of the structure limit, and removing the controller's ceiling bound leaves its three runs unchanged.

Table 3.7 Shipped Loops, Retuned Loops and Model Predictive Control on the Same Batches, Seed 20260918, Each Run to the End of Its Batch

Structure	IAE (mm min)	Margin (mm/min)	Pull (mm/min)	Heater (kW)	Prime (kg)	End (min)
Scenario 3, setpoint step						
shipped loops	378.5	0.386	1.46	12.9	84.63	494.9
retuned loops	149.6	0.338	1.28	15.2	85.01	504.6
predictive	81.0	0.325	1.65	11.8	85.80	512.2
Scenarios 5 and 6, ceiling case						
shipped loops	252.0	0.358	35.88	106.0	88.00	492.8
retuned loops	74.3	0.451	27.77	101.4	88.00	508.8
predictive	71.1	0.403	18.02	56.4	88.02	519.1
Scenario 11, setpoint step and argon upset						
shipped loops	462.0	0.309	35.60	105.9	84.44	484.4
retuned loops	176.0	0.325	27.53	104.0	85.02	497.3
predictive	119.1	0.313	18.32	58.2	85.64	507.8

IAE is the scorecard figure on the measured diameter, margin the minimum of $V_{max} - V_p$ at every 0.05 min physics step, and pull and heater the travel of the logged commands. The shipped ceiling row is scenario 6. The controller made 1016 to 1039 plans per run and none was rejected.

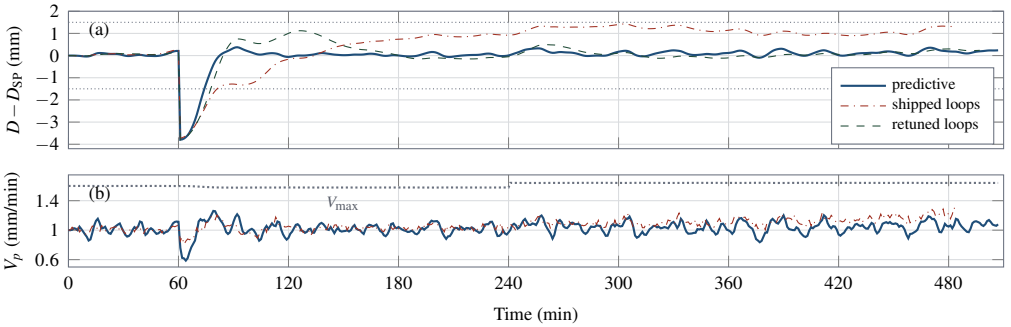


Figure 3.5 Scenario 11 under three structures to the end of each batch, with the setpoint step at 60 min and the argon upset at 240 min. The predictive run is the thick solid line, the shipped loops the dash-dotted line and the retuned loops the dashed line. (a) True diameter minus setpoint, with the tolerance as thin dotted lines. (b) Applied pull rate, with the structure limit of the predictive run as the thick gray dotted line.

✦ DID YOU KNOW? Better Tracking With Less Travel

From the simulator: on scenario 11 in Table 3.7 the predictive controller moves its actuators about half as far as the shipped loops, 18.32 mm/min of pull-rate travel against 35.60 and 58.2 kW of heater travel against 105.9, while it cuts the diameter IAE from 462.0 to 119.1 mm min. Better tracking need not cost more actuator movement.

The remainder, 176.0 against 119.1 mm min on scenario 11, belongs to the pull-rate plan, a model of the meniscus driven by the estimate of d. With the heater command frozen at its initial value inside the plan the IAE is still 130.5, while the margin falls to 0.136 mm/min and the prime crystal to 79.65 kg, which is what the heater buys, as under mid-ranging. With no process noise on d, and so no estimate, the IAE is 1700.4. The controller cuts the pull rate to 0.59 mm/min within four minutes of the step, and at 80 min the true diameter is 303.74, 303.08 and 302.41 mm under the controller, the retuned loops and the shipped loops, which reach 303.5 mm at 78.60, 81.85 and 111.35 min, as shown in Figure 3.5. From 250 to 300 min the shipped loops ride 1.14 to 1.40 mm high, because a 120 min integral time never catches the drain, against −0.02 to 0.33 under the controller and 0.05 to 0.49 under the retuned loops.

Table 3.8 Graded Challenge as Supplied and With the Suggested Configuration. Both Runs Carry the Same Disturbances, the Same Timed Events and the Same Seed, and Every Figure Is Read at the End of the Batch. Max $|\Delta D|$ Is on the Measured Diameter, While the Challenge Rule Judges the True Diameter.

	As supplied	Suggested
Architecture	DIC-301 only	DIC-301 + VIC-301
Schedule of Eq. (3.14)	off	on
K_c DIC-301 (mm/min per mm)	0.012	0.055
τ_I DIC-301 (min)	120	90
τ_D DIC-301 (min)	8	45
K_c VIC-301 (kW per mm/min)	2	8
τ_I VIC-301 (min)	600	120
Max $\|\Delta D\|$ (mm)	12.856	0.900
Final diameter (mm)	312.91	300.75
Worst margin (mm/min)	0.195	0.398
Prime crystal (kg)	12.80	88.02
Diameter IAE (mm min)	2313	189
Batch duration (min)	453.8	495.8
Value ($)	11,521	79,215

The controller does not win every measure. On the ceiling case the faster mid-ranging scenario 6 asks for (60 kW per mm/min, 10 min) keeps 0.454 mm/min of margin against 0.403. The margin of scenario 11 ties the shipped loops' only on the shipped noise sequence, and over ten other seeds it is 0.166 to 0.322 mm/min, below the shipped loops' 0.292 to 0.360 on all ten and the retuned loops' 0.208 to 0.308 on seven, while its IAE of 114.6 to 119.3 beats the retuned loops' 172.3 to 179.0 on all ten. Over nine cases (adding scenarios 4, 7, 8, 9, 10 and the challenge) the controller has the lower IAE on eight and the retuned loops the larger margin on eight, the challenge tying at 0.509 against 0.508. In the tail of scenario 8 it keeps 7.73 kg prime of 18 against 15.86 for the retuned loops, because its price draws the pull rate toward the 1.35 mm/min target above the defect band (all 18.00 kg with $r_V = 0$), as Section 3.6.3 found for mid-ranging.

Scenario 9 is the case the controller loses on IAE, 178.1 against 151.3 mm min for the retuned loops. As loaded, the shipped loops lose the crystal on all eleven noise seeds tested and the controller keeps it on all eleven, but on only six without its ceiling bound. Given $\hat{V}_{\max} - 0.02\,\mathrm{mm/min}$ as the output limit of DIC-301, the shipped loops also keep it on all eleven, so the crystal is saved by that estimate, which the loops can carry as a clamp, and not by the plan.

3.6.5 The Graded Challenge

The challenge scenario applies a hot-zone upset at $t = 90\,\mathrm{min}$ and a crucible lift drift at $t = 300\,\mathrm{min}$ on a hot zone already leaking 8 %, and asks for the diameter held within $\pm 1.5\,\mathrm{mm}$ of 300 mm for the whole batch with the pull rate still clear of the structure limit at the end. It opens with a soft diameter loop, a short derivative time, camera noise of 0.20 mm and the mid-ranging loop switched off, with a detuned mid-ranging setting waiting behind it. Table 3.8 reports both configurations.

The supplied tuning does not oscillate. It simply never catches up, and the crystal finishes 12.91 mm oversize with a maximum deviation of 12.856 mm and 12.80 kg of prime crystal out of 88.02 kg grown. Evaluating the loop of Table 3.3 at the supplied $K_c = 0.012$ and $\tau_D = 8\,\mathrm{min}$ gives a phase margin of 4.6° at a crossover period of 154.6 min, so the derivative time is a twentieth of the closed-loop response rather than comparable to it, and the soft proportional gain is the only thing keeping the loop out of oscillation. The suggested configuration meets the target with a maximum deviation of 0.900 mm, a worst margin of 0.398 mm/min and the full 88.02 kg prime. Running the same suggested tuning with the schedule switched off gives a maximum deviation of 0.999 mm

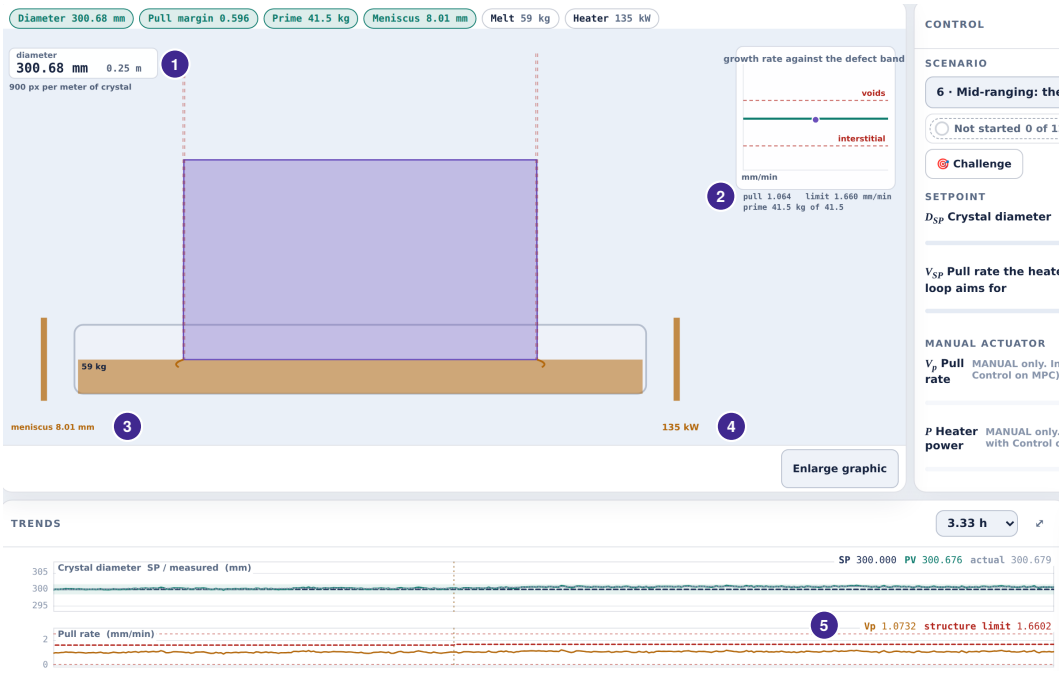


Figure 3.6 Scenario 6 running at $t = 4.00$ h, with the leaking hot zone and the argon upset already absorbed, cropped to the plant view and the first two trend strips. The diameter holds 300.68 mm against a 300.0 mm setpoint (1). The pull-rate command reads 1.073 mm/min on the second strip (5) against the structure limit drawn as a red dashed line at 1.660 mm/min, and the applied pull rate of 1.064 mm/min, printed with the limit under the growth-rate inset (2), leaves 0.596 mm/min of margin. The heater has been walked to 135.1 kW by the mid-ranging loop (4), the meniscus is at 8.01 mm between its flooding and detachment limits (3), and all 41.5 kg grown so far is prime (2).

at an essentially unchanged worst margin of 0.406 mm/min, so the schedule is worth 0.100 mm of diameter on this scenario and nothing at all in margin, which is what a loop-shape correction should be worth on a run that never approached the limit. The benefit is small but consistent, between 0.06 and 0.17 mm over nine noise seeds and positive on every one.

The challenge rule itself judges the true diameter at every logged sample and the margin at the end of the batch. Judged that way, the supplied settings end the batch at 453.8 min with a worst diameter error of 12.91 mm, more than eight times the band, and 0.195 mm/min of margin, so the attempt fails. The suggested settings end it at 495.8 min with a worst error of 0.75 mm, inside the band with 0.75 mm to spare, and 0.439 mm/min of margin.

Figure 3.6 shows scenario 6 four hours into the batch. The strip that carries the argument is the second one, where the pull rate is plotted against the moving structure limit, because the distance between those two traces is the quantity the mid-ranging loop exists to protect and the one the diameter trend above it cannot show.

3.7 Technology in Practice

The loops of Section 3.5 run on a camera, a pull drive, a heater supply and a controller, drawn together in Figure 3.7.

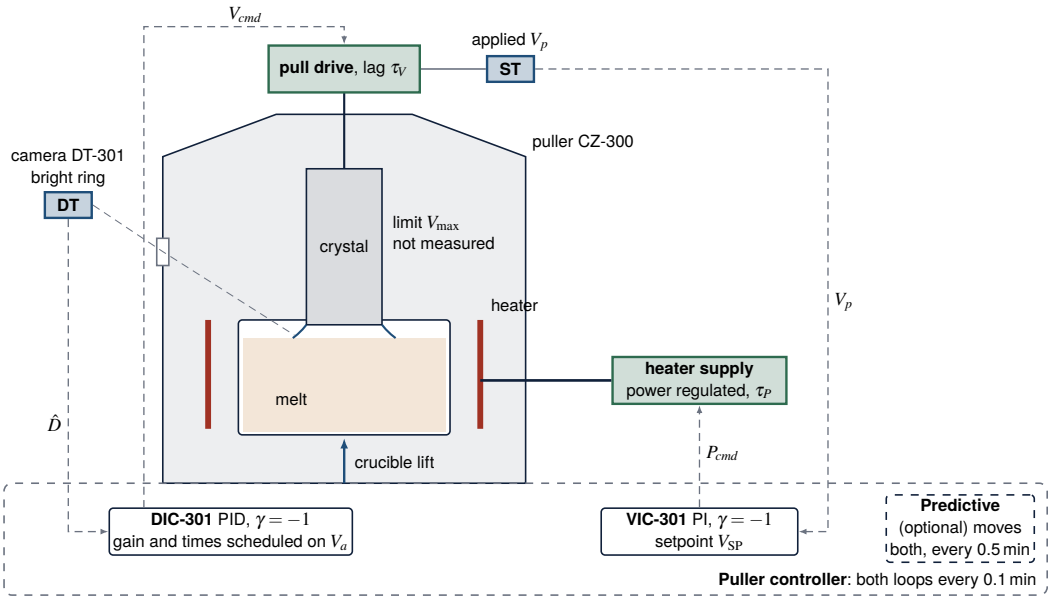


Figure 3.7 The loops of puller CZ-300 drawn with their hardware. Camera DT-301 sights the bright ring through a viewport, and DIC-301 sets the pull drive's command with its gain and times scheduled on the applied pull rate. VIC-301 reads that applied pull rate and sets the power of the heater supply. Both loops are direct acting ($\gamma = -1$). The optional predictive controller replaces both when it is selected. Notice that the structure limit has no instrument: VIC-301 protects it only by keeping the pull rate near V_{SP}.

What Measures It

The diameter loop reads camera DT-301, which the model reduces to noise of 0.35 mm by default and a filter of 0.50 min (Equation (3.10)).

> **TECHNOLOGY** **Sensor: A Camera on the Bright Ring**
>
> The bright ring is the reflection of the hot crucible wall in the curved meniscus where the crystal meets the melt. Optical pyrometers, photocells, line-scan cameras and two-dimensional cameras have all been used to find it. A camera in a viewport sees the ring at an angle, as an ellipse, and on a large crystal the crystal itself can hide the ring's widest point. Such instruments need frequent calibration by the operator [22]. A crystal that swings in a small orbit moves the ring, and the camera reports a diameter change that is not there. The loop then answers it with pull rate or heater and loses true control of the diameter [23]. A slower failure is a viewing window clouded by deposits, which weakens the signal a pyrometer receives through it [17].

The heater loop reads the pull rate the drive applies. The 1972 design weighed the crystal instead and controlled the ratio of weight to length, which is proportional to its cross-section [17], the weighing route of Section 3.2.

What Moves It

The pull drive follows its command through a lag of 0.30 min and is clamped at its mechanical limit of 2.50 mm/min, above the structure limit. A crystal growing fat is answered by pulling faster, so DIC-301 is direct acting, $\gamma = -1$ in the convention of Section 1.6.2.

TECHNOLOGY Final Element: The Pull Drive

The crystal hangs from a pull wire on a lifting mechanism at the top of the puller. The mechanism lowers a small seed into the melt and then withdraws it slowly, turning the crystal one way while the crucible turns the other [22]. A second drive raises the crucible. In the 1972 design one controller output set both drives through a distributor that kept the ratio of pull rate to crucible lift constant [17]. Because that loop needed strong derivative action, a step in its measurement became a jump in pull rate: a load cell with too large a dead band made the pull rate change abruptly and spoiled the resistivity along the crystal [17]. On a trend it is the pattern of Table 3.6, a busy pull rate under a calm diameter.

The heater supply follows its command through a lag of 0.50 min. The 1972 design already regulated power rather than temperature: an inner loop set the firing angle of thyristors from a measurement of the heater's power, which caught mains voltage swings before they warmed the heater and cut power fluctuations to a hundredth or less [17]. VIC-301 is direct acting as well, because a pull rate above its target calls for more heat.

Where the Controller Runs

Nested loops, in which one loop's setpoint is based on another loop's process variable, run with cycle times from milliseconds to minutes, and a programmable controller can serve as the primary controller of a smaller system [24]. Industrial pullers have used cascades of PID controllers [2]. The app executes both loops every 0.1 min against a physics step of 0.05 min, short beside the crossover period of 36.3 min. In a control system the structure is two standard blocks: a PID block whose gain and times the schedule multiplies, and a PI block whose process variable is the first block's actuator. The predictive controller of Section 3.5.3, which decides every 0.5 min, replaces both blocks with one plan, and the two blocks take over from the next scan whenever a plan is rejected.

IN PRACTICE What a Production Puller Controls

A study of an industrial 300 mm puller lists four controlled variables, the crystal radius, the growth rate, the heater temperature and the position of the melt surface, and three manipulated ones, the pull rate, the crucible lift rate and the heater power. Cascade control with PID controllers has been the industrial practice. The heater acts on the controlled variables with a long time constant, and the relation changes as the melt drains [2]. The diameter camera needs frequent calibration by the operator to keep the crystal within specification [22], and a camera bumped after calibration reads a wrong diameter [23]. The two loops of this chapter are the pull-rate and heater part of that cascade.

What Is Modernizing It

Three developments bear on this machine. The first is vision: two cameras on opposite sides of the crystal tell a swinging crystal, a moving melt level and a bumped camera from a real change of diameter [23]. The second is models trained on plant data. A gray-box model of a 300 mm puller, built toward predictive control, learns the hidden interface gradient from production records [2]. The third changes the hot zone. Magnetic fields that suppress melt convection and continuous feeding of liquid silicon were both listed as major developments in 1985, the second then far from a reliable production process [1]. A continuous feed would hold the melt inventory and remove much of the drift that makes this batch plant non-stationary.

3.8 Limitations

The model is not intended for hot-zone design. There is no radiation view factor, no heat shield geometry, no crystal or crucible rotation and no magnetic field, so the melt flow that dominates a real puller is absent and the thermal chain is three lumped nodes standing in for it [13]. The meniscus is the linearized form rather than the capillary solution, which restricts the model to body growth and places the flooding and detachment heights by hand. There is no axial gradient field, so the

defect criterion of (3.6) is applied to the growth rate alone and would have to be reformulated on V/G before a hot-zone change could be studied [5]. Oxygen, carbon and dopant transport are absent entirely, which removes resistivity control, striations and axial segregation, all of which are real control problems on the same machine.

Two limitations concern the control side specifically. The structure limit is a latch with a dwell rather than a progressive breakdown of the interface, so the model offers partial credit only in the sense that a short excursion is survivable, and it gives no early warning in the ring image at all. The dwell t_{lose} is one constant standing in for dislocation nucleation kinetics, and a real interface is more forgiving of a brief excursion at small diameter than of the same excursion at 300 mm, which the model does not represent. Exact repeatability is itself a fiction, since a frozen step and a frozen scan give every machine the same run and no furnace behaves that way. Three features came through every simplification, the double integrator through the meniscus, the unmeasured process constraint on the fast actuator and the non-stationary batch, and each one is present because it changes what the controller has to do. Section 3.3.7 says where each individual assumption gives out.

The predictive controller's model is linearized once at the full charge and the design hot zone, so the draining melt, a changed hot zone, a lift error and any edited constant reach it only through d. Its meniscus band rests on an estimate off by up to 0.65 mm in scenario 11 and 1.97 mm in scenario 9, the tangent of (3.2) and the latches are absent, and the weights were tuned once, on scenario 11.

> ⚠ **OPEN CHALLENGE** **The Gradient Nobody Measures**
>
> A predictive controller for a production puller needs a model of the whole batch, and most published control models were built for crystals under 200 mm. The temperature gradient in the crystal at the interface cannot be measured, and a first-principle model that held it constant fell short. A model that learned the gradient from crystal length and melt temperature in plant data cut the errors most in the late phase of the batch, where the gradient changes fastest [2]. The same gradient decides the defect band of Section 3.6.3. Progress would be a gradient estimate checked during the batch, so that the defect specification could be held as a ratio and not through the growth rate alone.

3.9 Discussion

The economics behind Table 3.5 are not subtle. A run occupies a furnace for eight hours, consumes a hundred kilograms of polysilicon and the electricity to melt and hold it, and the difference between the two scenario 9 rows in Table 3.5 is 2.84 kg of prime crystal and, with the shipped noise sequence, the single-crystal structure of everything grown after $t = 483.0$ min. Czochralski growth supplies essentially all of the single-crystal silicon used for semiconductors and a large share of that used for photovoltaics, where the energy consumed in crystal growth and wafering is one of the larger contributions to the energy payback time of a finished module [25]. A control structure that converts a lost run into a completed one is therefore an energy measure as well as a yield measure.

The equations given here also carry methods beyond the two loops. The ten-state system with an explicit constraint on a manipulated variable is a natural model predictive control problem, because $V_{\max}$ of (3.4) can enter as a state-dependent output constraint instead of being defended by a second loop, and the heater and the pull rate can be coordinated as one two-input problem rather than two, as Section 3.5.3 does in linear form. The unmeasured crucible lift error of scenario 7 is a textbook target for moving horizon estimation on the same equations, since it enters (3.1) where the diameter and the pull rate together carry enough information to separate it from a hot-zone change. Model-based designs of both kinds have been reported on real pullers [9–11], including the observer that reconstructs the radius from a weighing signal and the scheduling of controller parameters along the batch. Mid-ranging remains the useful comparison for all of them, because it is the answer available without a model and the one an engineer can commission in an afternoon [3, 4].

3.10 Conclusion

This chapter has used a Czochralski puller as an interactive laboratory for the case where the fast actuator is also the limited one. Ten states are integrated, the crystal radius and the meniscus height, a three-node thermal chain from heater to crucible to melt, the melt inventory, the crystal length, two actuator lags and the prime crystal grown so far, with the diameter reaching the pull rate only through the meniscus and therefore through two integrations. Without derivative action the loop has a phase margin of −6.2°, the diameter overshoots a 4 mm setpoint step by 5.8 mm, and the meniscus floods at 158.4 min, while 45 min of derivative time gives 29.2° of margin against a crossover period of 36.3 min. Adding a proportional-integral loop that drives the heater from the pull rate rather than from the diameter raises the worst pull-rate margin on the same disturbances from 0.221 to 0.358 mm/min and the prime crystal from 73.46 to 88.00 kg, and on a degraded bright-ring camera it is the difference between losing structure at 483.0 min and completing the batch with the shipped noise sequence. Both the gain and the natural frequency of the diameter plant are proportional to the pull rate, so scheduling K_c upward and τ_D, τ_I and τ_F downward in that variable holds the phase margin of the meniscus loop at 29.19° across the whole drive range, where a fixed tuning slides from 38.1 to 21.3° and scaling the gain and the derivative time the other way is worse than no schedule at all below the design pull rate. The camera filter and the drive lag are inside the loop and outside the schedule, and they are what leaves 25.2 to 13.9° in place of a flat curve on the complete loop. An optional predictive controller lowers the diameter IAE of scenario 11 from 462.0 to 119.1 mm min, and loops retuned by the same rule reach 176.0 with more margin. Further work would add a capillary meniscus, an axial gradient field so that the defect criterion becomes a ratio, melt convection as a disturbance with structure, and the shoulder as an open-loop trajectory problem. The Czochralski puller is an accessible browser platform for learning that an actuator limit is a control problem for a different loop from the one using the actuator.

KEY POINTS

- The pull rate reaches the diameter only through a short column of melt, two integrations away.
- Without derivative action the phase margin is −6.2° and the meniscus floods. A 45 min derivative time gives 29.2°.
- The diameter loop holds the diameter while the pull rate walks toward a structure limit that nothing measures.
- Driving the heater from the pull rate raises the worst margin from 0.221 to 0.358 mm/min and the prime crystal from 73.46 to 88.00 kg.
- A schedule on the pull rate holds the meniscus loop at 29.19°, but the camera and drive lags leave 25.2 to 13.9°.

Where else this applies. Mid-ranging a fast actuator with a slow one in Volume 4, Chapter 10 and Volume 4, Chapter 9, and a gain schedule along a batch in Chapter 2.

References

[1] J. H. Matlock, “Future application of Czochralski crystal pulling for silicon,” in *Proceedings of the Flat-Plate Solar Array Project Workshop*. Jet Propulsion Laboratory, 1985, NASA Technical Reports Server 19860010255.

[2] S. Kato, S. Kim, M. Kano, T. Fujiwara, and M. Mizuta, “Gray-box modeling of 300 mm diameter Czochralski single-crystal Si production process,” *Journal of Crystal Growth*, vol. 553, p. 125929, 2021.

[3] F. G. Shinskey, *Process Control Systems: Application, Design, and Tuning*, 4th ed. New York, NY: McGraw-Hill, 1996.

[4] B. J. Allison and A. J. Isaksson, “Design and performance of mid-ranging controllers,” *Journal of Process Control*, vol. 8, no. 5-6, pp. 469–474, 1998.

[5] V. V. Voronkov, “The mechanism of swirl defects formation in silicon,” *Journal of Crystal Growth*, vol. 59, no. 3, pp. 625–643, 1982.

[6] W. von Ammon, E. Dornberger, H. Oelkrug, and H. Weidner, "The dependence of bulk defects on the axial temperature gradient of silicon crystals during Czochralski growth," *Journal of Crystal Growth*, vol. 151, no. 3-4, pp. 273–277, 1995.

[7] D. T. J. Hurle, "Control of diameter in Czochralski and related crystal growth techniques," *Journal of Crystal Growth*, vol. 42, pp. 473–482, 1977.

[8] ——, "Analytical representation of the shape of the meniscus in Czochralski growth," *Journal of Crystal Growth*, vol. 63, no. 1, pp. 13–17, 1983.

[9] M. A. Gevelber and G. Stephanopoulos, "Dynamics and control of the Czochralski process," *Journal of Crystal Growth*, vol. 84, no. 4, pp. 647–668, 1987.

[10] J. Winkler, M. Neubert, and J. Rudolph, "Nonlinear model-based control of the Czochralski process I: Motivation, modeling and feedback controller design," *Journal of Crystal Growth*, vol. 312, no. 7, pp. 1005–1018, 2010.

[11] M. Neubert and J. Winkler, "Nonlinear model-based control of the Czochralski process III: Proper choice of manipulated variables and controller parameter scheduling," *Journal of Crystal Growth*, vol. 360, pp. 3–11, 2012.

[12] J. Czochralski, "Ein neues verfahren zur messung der kristallisationsgeschwindigkeit der metalle," *Zeitschrift für Physikalische Chemie*, vol. 92U, no. 1, pp. 219–221, 1918.

[13] W. Zulehner, "Czochralski growth of silicon," *Journal of Crystal Growth*, vol. 65, no. 1-3, pp. 189–213, 1983.

[14] K. J. Gärtner, K. F. Rittinghaus, A. Seeger, and W. Uelhoff, "An electronic device including a TV-system for controlling the crystal diameter during czochralski growth," *Journal of Crystal Growth*, vol. 13-14, pp. 619–623, 1972.

[15] W. Bardsley, G. W. Green, C. H. Holliday, and D. T. J. Hurle, "Automatic control of Czochralski crystal growth," *Journal of Crystal Growth*, vol. 16, no. 3, pp. 277–279, 1972.

[16] W. Bardsley, D. T. J. Hurle, and G. C. Joyce, "The weighing method of automatic Czochralski crystal growth," *Journal of Crystal Growth*, vol. 40, no. 1, pp. 13–20, 1977.

[17] T. Kimoto, A. Wachi, S. Sakurai, and M. Mikami, "Automatic diameter control of Czochralski single crystal rods," *Transactions of the Society of Instrument and Control Engineers*, vol. 9, no. 5, pp. 595–600, 1973, in Japanese, with an English abstract.

[18] W. J. Rugh and J. S. Shamma, "Research on gain scheduling," *Automatica*, vol. 36, no. 10, pp. 1401–1425, 2000.

[19] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, "Constrained model predictive control: Stability and optimality," *Automatica*, vol. 36, no. 6, pp. 789–814, 2000.

[20] S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Control Engineering Practice*, vol. 11, no. 7, pp. 733–764, 2003.

[21] G. Pannocchia and J. B. Rawlings, "Disturbance models for offset-free model-predictive control," *AIChE Journal*, vol. 49, no. 2, pp. 426–437, 2003.

[22] R. H. Fuerhoff, "Non-distorting video camera for use with a system for controlling growth of a silicon crystal," U.S. Patent 5,656,078, filed 14 November 1995, 1997, granted 12 August 1997.

[23] A. LaBrie and M. Baba, "Two camera diameter control system with diameter tracking for silicon ingot growth," U.S. Patent 6,030,451, filed 12 January 1998, 2000, granted 29 February 2000.

[24] K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, A. Hahn, S. Saravia, A. Sherule, and M. Thompson, "Guide to operational technology (OT) security," National Institute of Standards and Technology, Gaithersburg, MD, Special Publication 800-82r3, 2023.

[25] T. Wetzel and S. Borchers, "Update of energy payback time and greenhouse gas emission data for crystalline silicon photovoltaic modules," *Progress in Photovoltaics: Research and Applications*, vol. 23, no. 10, pp. 1429–1435, 2015.

3.11 Exercises

3.1 ○ Concept check · 2 min

Scenario 3 is run with the diameter derivative time set to zero. The setpoint steps from 300 to 304 mm at 60 min, the camera noise is off, and the run stops itself at 158.4 min. What ends it?

(a) The pull rate crosses the structure limit and stays above it for the whole dwell

(b) The pull-rate command reaches the drive's mechanical limit and the integral winds up

(c) The growing oscillation drives the meniscus below its flooding height

(d) Camera noise differentiated without a filter spikes the pull rate and scraps the crystal

3.2 Concept check · 1 min

Which signal does mid-ranging controller VIC-301 take as its measurement?

(a) The diameter error, filtered more slowly than DIC-301 reads it

(b) The melt temperature, which the heater moves most directly

(c) The pull-rate command that DIC-301 sends to the pull drive

(d) The pull rate the drive actually applies, V_p

3.3 Concept check · 2 min

During a batch the applied pull rate falls from the design 1.0 mm/min to 0.5 mm/min. Which change to the settings of DIC-301 keeps the loop shape on the meniscus plant of Equation (3.12)?

(a) K_c halved, and τ_I, τ_D and τ_F each doubled

(b) K_c doubled, and τ_I, τ_D and τ_F each halved

(c) K_c doubled and τ_D doubled, with τ_I and τ_F unchanged

(d) K_c halved, with τ_I, τ_D and τ_F all unchanged

3.4 Theory and calculation · 6 min

Use $c_h = 0.050\,\mathrm{rad/mm}$, $r_0 = 150\,\mathrm{mm}$, $V_0 = 1.00\,\mathrm{mm/min}$ and the default DIC-301 settings $K_c = 0.055$ (mm/min per mm), $\tau_I = 120\,\mathrm{min}$, $\tau_D = 45\,\mathrm{min}$ and $\tau_F = 8\,\mathrm{min}$. The mid-ranging loop holds the pull rate at the 1.35 mm/min target of scenario 8.

(a) From Equation (3.12), find the natural frequency, the natural period and the magnitude of the steady-state gain of the diameter plant at this pull rate, to three significant figures. *Checkpoint:* 0.0247 rad/min, 255 min and 222 mm per mm/min.

(b) Find the four settings the schedule of Equation (3.14) applies at this pull rate. *Checkpoint:* $K_c = 0.0743$, $\tau_I = 88.9\,\mathrm{min}$, $\tau_D = 33.3\,\mathrm{min}$, $\tau_F = 5.93\,\mathrm{min}$.

(c) Name the action of DIC-301 in the convention of Section 1.6.2, with the value of γ, and say whether the pull rate changes it. *Checkpoint:* Direct acting, $\gamma = -1$, at every pull rate.

3.5 Theory and calculation · 7 min

Scenario 5 opens with a hot zone that leaks 12 % more heat than the recipe assumed ($\lambda = 1.12$) and the rest of the design point unchanged: $r = 150\,\mathrm{mm}$, $V_g = 1.000\,\mathrm{mm/min}$, $T_m = 1425\,^\circ\mathrm{C}$, $m = 100\,\mathrm{kg}$. Use the constants of Table 3.2.

(a) Find the solidification rate and the latent heat it draws from the melt, in kW. *Checkpoint:* 0.1647 kg/min and 4.94 kW.

(b) Work Equation (3.11) up the chain to the crucible temperature, the heater temperature and the heater power, each to one decimal. *Checkpoint:* $T_c = 1582.8\,^\circ\mathrm{C}$, $T_h = 1727.9\,^\circ\mathrm{C}$, $P_a = 133.8\,\mathrm{kW}$.

(c) Compare the power with the design value at $\lambda = 1$ and say which term of the chain carries most of the difference. *Checkpoint:* 15.6 kW more, most of it the melt surface loss.

3.6 Theory and calculation · 6 min

At the design point a 1 kW change of heater power moves the steady melt temperature by 2.819 °C, and $K_T = 0.025$ (mm/min) per °C. The diameter loop crosses over with a period of 36.3 min.

(a) Find the static gain from heater power to the pull rate that the diameter loop settles at, and give the action of VIC-301 in the convention of Section 1.6.2. *Checkpoint:* −0.0705 mm/min per kW, direct acting.

(b) With the integral dominant, the mid-ranging loop returns the pull rate with a time constant near $\tau_I/(K_c|K_{PV}|)$, where K_{PV} is the gain of part (a). Find it and its ratio to the crossover period for

the shipped tuning ($K_c = 8$, $\tau_I = 120\,\text{min}$) and the retuned one ($K_c = 10$, $\tau_I = 55\,\text{min}$), with K_c in kW per mm/min. *Checkpoint:* 213 min and 5.9; 78 min and 2.1.

(c) Repeat for the challenge as supplied ($K_c = 2$, $\tau_I = 600\,\text{min}$) and compare with the length of a batch. *Checkpoint:* About 4260 min, some 71 hours, against a batch of about eight.

3.7 App challenge · 8 min

Load scenario 3, "A setpoint step, and why there is a derivative term" (default seed), and set Diameter derivative time to 0 before pressing Run. Run at the default 20 min/s until the run stops itself, then press Export CSV. The time column of the file is Time (h).

(a) Read the clock when the run stops, and the meniscus strip at that moment. What ended the run? Give the time to within 0.02 h. *Checkpoint:* 2.64 h; the meniscus has reached its flooding line.

(b) From the column `dia_mm` (the true diameter), find the highest value and the time of that row, to within 0.1 mm and 0.02 h. *Checkpoint:* 309.8 mm at 1.70 h.

(c) From the column `v_head_mm_min`, find the smallest pull-rate margin of the run to within 0.005 mm/min. Does it support (a)? *Checkpoint:* 0.214 mm/min: the structure limit was never reached.

3.8 App challenge · 8 min

Load scenario 5, "The limit the pull rate walks into" (default seed), and run it as loaded at 60 min/s until the run stops itself at the end of the batch. Read the Prime crystal tile and the clock, and export the CSV. Then press Reset, set Control to Diameter + mid-ranging, run to the end again, read the same two values and export a second CSV.

(a) Give the prime crystal and the end time of each run, to 0.1 kg and 0.02 h. *Checkpoint:* 73.5 kg at 7.75 h; 88.0 kg at 8.21 h.

(b) Find the smallest value of `v_head_mm_min` in each file, to within 0.003 mm/min. *Checkpoint:* 0.239 mm/min and 0.377 mm/min.

(c) Table 3.5 reports worst margins of 0.221 and 0.358 mm/min for the same pair. Explain the difference and say which number an engineer should trust. *Checkpoint:* The file logs two samples per minute, the table every physics step.

3.9 App challenge · 8 min

Load scenario 9, "A noisy bright ring against a long derivative time" (default seed), and run it as loaded at 20 min/s until it stops itself. Read the clock and the Prime crystal tile. Then switch the panel to Advanced, press Reset, set Derivative filter to 12.0 min, run at 60 min/s to the end of the batch, read the same two values and export the CSV.

(a) Give the stop time and prime crystal of the first run to 0.02 h and 0.1 kg, and the noise gain τ_D/τ_F of each run. *Checkpoint:* 1.89 h, 19.0 kg; 11.25 and 3.75.

(b) Give the end time and prime crystal of the second run, and the smallest `v_head_mm_min` in its file to within 0.005 mm/min. *Checkpoint:* 8.43 h, 88.0 kg, 0.086 mm/min.

(c) What did the longer filter cost, by the chapter's analysis? *Checkpoint:* 23.0° of phase margin, from 44.3 to 21.4°.

3.10 Application area · 3 min

A puller vendor proposes to replace the bright-ring camera with a load cell on the crystal, arguing that a weight is a cleaner signal than an image. From the chapter's history and model, what must a weighing measurement handle that the optical one does not, and what does the optical route cost the diameter loop of this chapter?

3.11 Application area · 4 min

(a) Prime crystal is valued at $900 per kg in the application. Using Table 3.5, what is one boule's difference in value between scenario 5 and scenario 6, and what else does the run cost that the valuation leaves out? *Checkpoint:* About $13,100 per boule, plus furnace time, polysilicon and energy.

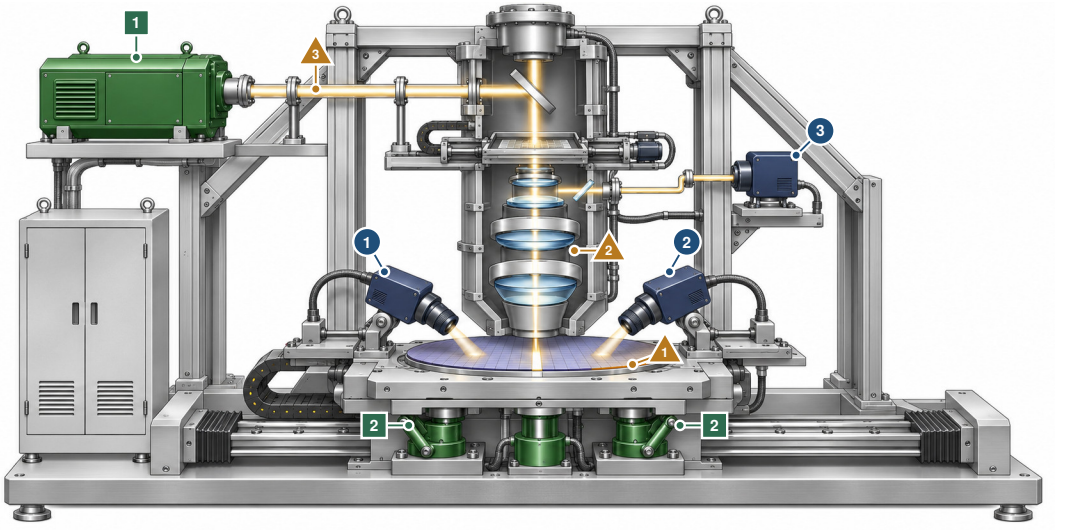
(b) In scenario 7 with mid-ranging, the diameter stays in band all run and the pull rate is walked back toward its target, while the heater ends at 121.59 kW having started at 118.26 kW. What should a process engineer conclude? *Checkpoint:* Something upstream moved, here the crucible lift.

3.12 Review the agent · 5 min

An AI assistant reviews scenario 9 and writes: "The 8 min derivative filter costs 21.7° of phase margin against 3 min, and the loop has little margin to spare. Set Derivative filter to 3.0 min. The margin rises to about 51°, the loop becomes better damped and the pull rate stays further from the structure limit." Find the flaw, and say how you would check it in the application.

CHAPTER 4

Wafer Scanner Simulator for Preview Feedforward Where No Measurement Exists



The controller moves 1 laser, whose pulse energy the dose loop sets
2 stage drives that lift and tilt the wafer, commanded from the sensor ahead

The controller reads 1 height sensor reading the surface ahead of the slit
2 height sensor reading the gap behind the slit, for the slow trim 3 energy monitor on the beam, read by the dose loop

What disturbs it 1 a bowed and rough wafer surface 2 a lens that warms and shifts the image size
3 laser pulses that scatter in energy

Notice that both height sensors aim beside the slit and never into it: the height where the light lands is never measured. Illustration, not to scale.

IN PLAIN TERMS **Keeping a Moving Wafer in Focus**

The main idea. Chips are printed by shining light through a narrow slot onto a silicon wafer that slides past underneath. The picture is sharp only within a very thin band of height. A stage therefore lifts and tilts the wafer as its uneven surface goes by.

The challenge. Nothing can measure the height where the light falls. A sensor reads the surface a short way ahead, and the stage moves early. How far ahead depends on how fast the wafer slides, so a faster machine quietly blurs the print.

Why it matters. Faster scanning prints more wafers each hour, and nothing on the machine warns that the focus has paid for it.

THREE WAYS TO READ THIS CHAPTER

Explore	11 min	The first page, the fact box in Section 4.1, the history box in Section 4.2, the fact and try-it boxes in Section 4.6, and the key points
Study	55 min	The whole chapter with its pause points and the twelve exercises
Practice	20 min	The first page, Section 4.6, Section 4.7, Section 4.8 with its open challenge, and the second challenge in Section 4.9

The control question Nothing can measure focus under the exposing slit of a scanner, so the stage is commanded from a level sensor that reads the wafer surface ahead of the slit. How far ahead should it read, why does that distance move whenever the scan speed is raised for throughput, and what is left for feedback to do?

Learning objectives

- Compute the group delay of a second-order stage servo and the matched preview distance at a given scan speed and bandwidth.
- Explain why a loop fed from a sensor behind the slit cannot be tuned into the depth of focus, from its transport delay, its servo lag and the frequency of the topography.
- Trace the offset of the swept optimum preview to the slope that the trailing-sensor trim reads, and set the trim slower than one field scan.
- Compute the dose floor $3\sigma_p/\sqrt{N}$ from the pulse count, and recognize a loop that has reached a statistical floor.
- Trade scan speed and alignment marks against focus, dose, overlay and wafers per hour.

Prerequisites Sections 1.4.7 (preview feedforward), 1.4.2 (what a tool can measure), 1.6.2 (velocity-form PID and loop action), 1.6.1 and 1.6.4 (runs, seed, tiles and data).

The app https://apmonitor.com/apps/litho-scanner/

See also Volume 3, Chapter 6 for feedforward from a measured disturbance with the late measurement demoted to a slow trim. Volume 5, Chapter 2 for control through a transport delay and Volume 5, Chapter 5 for qualities fixed before they can be measured. Volume 8, Chapter 10 for iterative learning on a repeated scan path and Volume 8, Chapter 9 for corrections from sampled, late results.

4.1 Introduction

A projection scanner exposes a wafer through a narrow slit, and the slit is where the light is. No sensor can occupy that space, so the one quantity the focus loop exists to regulate, the distance between the resist surface and the focal plane at the point being printed, has no instrument attached to it and never will. The measurement that does exist reads the surface a millimeter or two up the scan path, before the slit arrives. The stage is commanded from that reading. This is feedforward from a measurement of the disturbance itself, and the lead it supplies is set by geometry and by scan speed rather than by a tuning constant [1, 2]. Section 1.4.7 places this preview among the other feedforward paths of the fab.

Plants of this shape are common once the pattern is recognized. Any process that transports material past a fixed working point can measure the material before it arrives, and a controller that waits for an error at the working point is then accepting a delay it did not have to accept. Active suspensions driven by a road sensor looking ahead of the wheel are the textbook case [3], and the quarter car of Volume 7, Chapter 11 carries road-preview feedforward of this kind. The optimal-preview formulation that describes them is written for any plant whose future reference or disturbance is known a fixed interval in advance [4], including the case where the preview is wrapped around a feedback loop that is already commissioned [2]. The common feature is that the correct lead time is a property of the plant being driven, not of the disturbance, and that the correct lead *distance* is that time multiplied by the transport velocity. Change the velocity for throughput and a control parameter has moved with no indication anywhere that it did.

A student meets feedforward as a block diagram in which a disturbance transfer function is divided by a process transfer function, and the lesson usually stops at the algebra. Two things that decide whether the scheme works are missing from that picture. The first is that the lead is physical, so it is installed as a distance on a machine and is right only at one speed. The second is that the feedforward path is blind to any error in the sensor it reads, which is why a preview controller needs a slow feedback trim beside it and why that trim is a different loop with a different purpose. Neither point survives a transfer-function derivation, and both are visible in half a minute on a plant that exposes a wafer.

Simulation of lithography has been developed principally on the optical and chemical side, where aerial image, resist and process window models answer questions about linewidth [5, 6]. The mechatronic side is described in the control literature mostly at the level of a single stage axis, its resonances and its feedforward identification [1]. However, the machine-level statement that connects the two, that the wafer per hour number and the focus budget are set by one shared parameter, is difficult to demonstrate without a tool that writes a whole wafer and reports what came out of it. This chapter works with an interactive step-and-scan exposure simulator in which the focus loop is commanded from a sensor that reads ahead of the slit, so that the matched preview distance, the dose noise floor and the throughput can be driven from one panel and their coupling read off a finished wafer. Preview control and scanner physics are established subjects. What the chapter measures is how far the standard group-delay estimate of the matched preview distance sits from the swept optimum, and why.

> ✦ **DID YOU KNOW? A Heavy Stage Placed to Nanometers**
>
> A wafer scanner moves masses of around 100 kg, accelerates them at over 100 m/s^2, and places them to within nanometers in every degree of freedom in a few milliseconds [7]. A modern wafer may pass through the lithographic cycle up to 60 times [7]. The height axis of this chapter's stage is limited to 30 m/s^2.

Section 4.2 traces how the scanning slit created the focus control problem. Section 4.3 gives the eight-state model of the stage, the laser, the lens and the exposure sequence, with the hundred-field layout, the six-parameter grid fit and the step-and-scan state machine, in enough detail to rebuild the plant, and states where each assumption fails. Section 4.4 derives the matched preview distance from the servo group delay and compares it with a swept optimum. Section 4.5 describes the four corrections: preview feedforward with a trailing trim, open-loop leveling, the dose loop and the alignment fit. Section 4.6 runs the scenarios in an order that pulls apart architecture from tuning, spatial frequency from amplitude, and a statistical floor from a loop that is merely detuned, and ends with the graded challenge. Section 4.7 names the sensors, the stage and the controller behind the corrections. Sections 4.8 to 4.10 state the limitations, discuss the throughput trade and conclude.

4.2 From Contact Printing to the Scanning Slit

Early integrated circuits were printed by pressing a mask against the wafer and exposing the whole substrate at once. Contact printing gave excellent resolution and destroyed masks, proximity printing spared the mask and lost resolution to the gap, and neither had a focus control problem worth the name because the mask and the wafer moved together. Projection printing separated them and introduced the optical requirement that the resist sit within the depth of focus of a lens. Reduction step-and-repeat projection printers arrived in the late 1970s and dominated production for fifteen years, exposing one field, stepping and exposing the next [8].

A stepper can level. The field is stationary while the shutter is open, so the tool measures the surface under the stationary field, drives the stage to best focus, and only then exposes. Field-by-field leveling in exactly that form was introduced on high numerical aperture i-line steppers in the late 1980s [9] and was combined with a global alignment grid shortly afterward [10]. Both papers describe a measurement that finishes before the exposure starts, which is the defining property of the stepper era and the reason focus was a positioning problem rather than a tracking problem.

Step-and-scan began to replace step-and-repeat during the 1990s because the field outgrew what a lens could image at once. Scanning a slit across the reticle and the wafer together buys a large field from a small well-corrected image area, at the price of having to synchronize two stages while the light is on [11]. That change created the focus control problem in its present form. A scanning tool cannot level a field in advance, because the point being exposed moves across the field at hundreds of millimeters per second, so the stage must follow the surface during the exposure. Level sensing that reads ahead of the slit, and the software delay that matches the lead to the stage servo, arrived with

the architecture. The resulting focal plane budget was measured and published early in the scanner era [12], and the incoming wafer specification that supports it moved from a whole-wafer flatness number toward a site flatness number defined over one field, because a scanner cares about height change per unit of travel rather than about total thickness variation [13].

Two of the other three corrections in a scanner have equally specific origins. Exposure dose on an excimer-illuminated tool is delivered as a burst of discrete pulses, and the statistics of that burst were analyzed as a control problem as soon as excimer sources were proposed for lithography [14]. The process sensitivity those pulses feed, the dependence of linewidth on dose and on focus, is the Bossung description of the process window that has been standard since 1977 [5, 6]. Lens heating, the slow absorption of exposure energy in the projection optics and the magnification and focus drift that follows it, was characterized once tools became productive enough for the thermal transient to matter within a single lot [15]. All three of these survived the move to immersion and then to extreme ultraviolet, where the numbers tightened but the structure did not [16].

> ⧖ HISTORY **From Three Points to Every Field (1982 to 1991)**
>
> A patent application filed in 1982 described how a reduction projection printer leveled a wafer: an autofocus read the surface at three points and the whole wafer was set square to the lens. Larger wafers, new substrate materials and every round of exposure and processing made the surface less even, so each part of the wafer needed its own height, read with light sent in at an oblique angle [17]. In 1987 the autofocus of a stepper was turned into a gauge of flatness, and on new wafers it found, in the worst case, more than four micrometers of thickness variation within one field [13]. By 1991 a stepper leveled every field and still wrote more than 80 wafers of 150 mm an hour, because it no longer aligned each field separately [10]. The level sensor of this chapter reads the surface the same way, but while the wafer moves.

4.3 Process Model

The simulated tool writes a hundred 22 mm by 24 mm fields on a 300 mm wafer. Field centers sit on a 22 mm by 24 mm pitch and are selected in order of distance from the wafer center until one hundred are collected, then reordered into the serpentine the stage actually drives. The slit is 6 mm tall along the scan and a full field wide, so one field scan covers $L = H + h_s = 30\,\text{mm}$ of travel. Figure 4.1 shows a section through the slit along the scan direction, carrying every symbol the equations of this section use.

What the four corrections act on and act against is set out in Table 4.1. Working units are the ones a scanner engineer quotes, seconds for time, millimeters in the wafer plane, micrometers for stage height, microradians for tilt, nanometers for defocus, overlay and linewidth, and mJ/cm^2 for dose. Eight states are integrated. Two of the three controlled variables have no instrument on the machine, which is the subject of this chapter.

4.3.1 The Surface the Stage Has To Follow

The wafer surface is the sum of three terms of very different character, as shown in Equation (4.1).

$$z_w(x,y) = b\left(\frac{r^2}{R^2} - \frac{1}{2}\right) + a\,\sin\frac{2\pi x}{L_1}\cos\frac{2\pi y}{L_2} + \rho_t\,\varepsilon(x,y;L_m) \tag{4.1}$$

where b is the chuck clamp bow from center to edge, $r = \sqrt{x^2+y^2}$ is the radius in the wafer plane, R is the wafer radius, a is a long-wavelength shape of wavelengths L_1 and $L_2 = 0.74L_1$, ρ_t is the amplitude of the short-wavelength topography, and ε is a smooth pseudo-random field of unit amplitude and correlation length L_m built by hashing a lattice of pitch L_m and interpolating with a smoothstep. Every term is deterministic in (x,y), so the same wafer is the same wafer on every run and a sweep of a control parameter is a controlled experiment rather than a sample of a random surface.

The three terms carry the whole argument of this chapter. At 400 mm/s the bow of 1.5 µm across the 300 mm diameter presents itself to the servo at about 1.3 Hz, and over the hundred exposed fields

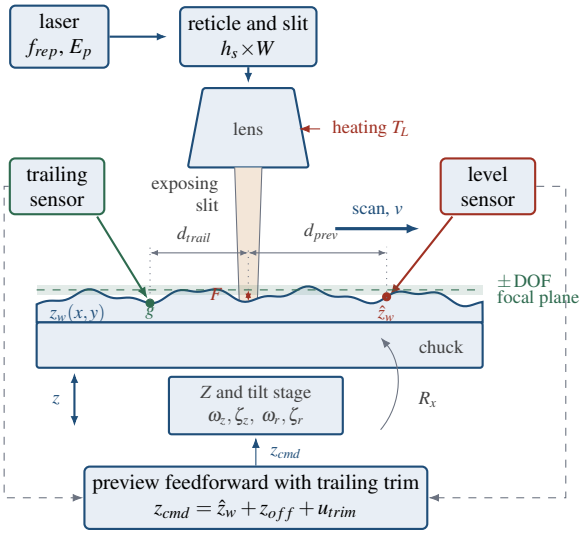


Figure 4.1 Section through the exposing slit along the scan direction. The exposure point travels left to right across the wafer at speed v, and the defocus F is the gap between the surface z_w and the focal plane at the slit, where no sensor can be placed because the slit is where the light is. The level sensor reads the surface d_{prev} ahead of the slit and supplies the feedforward command, the trailing sensor reads a gap g that is d_{trail}/v seconds old and supplies the trim, and the Z and tilt stage answers the command with natural frequencies ω_z and ω_r and damping ratios ζ_z and ζ_r. Only the incident ray of each oblique sensor is drawn, and the vertical scale of the surface is exaggerated by about four orders of magnitude against the horizontal.

the stage height command travels 1.74 µm and the stage itself 1.68 µm, while the servo's following error on it adds next to nothing to the error budget. The topography term is 0.05 µm, a thirtieth of that amplitude, and at $L_m = 7$ mm it presents itself at $v/L_m = 57.1$ Hz, which is where a 60 Hz servo lives. A spatial disturbance becomes a temporal one when it is divided by a velocity, and what a servo answers to is amplitude multiplied by the square of the resulting frequency.

The tilt stage removes the plane fitted across the slit and nothing finer, so the slope it is asked for is taken as a chord over $\pm h_c = 5.5$ mm rather than as a local derivative, as shown in Equation (4.2).

$$\theta(x,y) = 10^3 \, \frac{z_w(x+h_c,y) - z_w(x-h_c,y)}{2h_c} \tag{4.2}$$

where θ is in µrad and the factor 10^3 converts µm/mm to µrad. What no plane can follow is the curvature of the surface across the slit, given by Equation (4.3).

$$\kappa(x,y) = \frac{z_w(x+h_c,y) + z_w(x-h_c,y) - 2z_w(x,y)}{2} \tag{4.3}$$

Equation (4.3) is in µm and enters the focus budget with a weight of 0.35, which represents the root-mean-square of the second-order departure over the slit. Together with the across-slit tilt term of (4.9) it sets a floor that no servo bandwidth reaches. The curvature alone contributes 7.7 nm over the hundred fields of the nominal wafer, and the floor is measured at 8.8 nm in Section 4.6.

4.3.2 The Stages, the Laser and the Lens

The height stage is a second-order position servo specified as a bandwidth and a damping ratio, because those are the two numbers a stage engineer quotes and because the preview distance is

Table 4.1 Classification of Variables. Nominal Values Are Those of the Design Operating Point at Which Every Scenario Starts.

Symbol	Role	Nominal	Range
F defocus at slit	CV, unmeasured	0 nm	±60 nm DOF
$R_x - \theta$ tilt error	CV, unmeasured	0 µrad	±5 µrad band
$\dot{D}$ dose rate	CV, measured	400 mJ/(cm^2 s)	band 4
z_{cmd}	MV	surface	±60 µm
$R_{x,cmd}$	MV	slope	±400 µrad
E_{cmd} pulse energy	MV	0.440 mJ	0.02–1.10 mJ
v scan speed	operator input	400 mm/s	100–800 mm/s
n_m marks per wafer	operator input	20	4–40
D_{SP} dose	SP	30 mJ/cm^2	10–60 mJ/cm^2
z_{off} focus offset	SP	0 nm	±200 nm
clamp bow scale	DV, unmeasured	1×	0–3×
topography scale	DV, unmeasured	1×	0–3×
laser gain g_p	DV, unmeasured	1×	0.8–1.2×
monitor gain g_m	DV, unmeasured	1×	0.95–1.05×
level sensor bias δ	DV, unmeasured	0 nm	±150 nm
grid T_x, T_y	DV, sampled	8, −5 nm	±60 nm
grid m, ρ	DV, sampled	0.40 ppm, 1.20 µrad	±2, ±5
non-correctable A	DV, unmeasured	1.2 nm	0–15 nm
z, $\dot{z}$	states	surface, 0	±60 µm
R_x, $\dot{R}_x$	states	slope, 0	±400 µrad
D_{acc}	state	0 mJ/cm^2	≥ 0
E_p	state	0.440 mJ	0.02–1.10 mJ
T_L lens heating	state	0	0–1
ξ scan coordinate	state	0 mm	0–30 mm

written in the same two numbers. The differential equations are Equations (4.4) and (4.5).

$$\ddot{z} = \mathrm{sat}\left[\omega_z^2(z_{cmd} - z) - 2\zeta_z\omega_z\dot{z},\ \pm a_{max}\right], \quad \omega_z = 2\pi f_z \tag{4.4}$$

$$\ddot{R}_x = \omega_r^2(R_{x,cmd} - R_x) - 2\zeta_r\omega_r\dot{R}_x, \quad \omega_r = 2\pi f_r \tag{4.5}$$

where z is in µm, R_x is in µrad, and $a_{max} = 30\,\mathrm{m/s^2}$ is the acceleration limit of the height axis expressed in µm/s^2 inside the saturation. The height state is clipped to $\pm z_{lim}$ and the tilt state to $\pm R_{lim}$ after every step, so a command that exceeds the stroke moves the axis to the stroke and no further. The laser follows its energy command through one lag, as shown in Equation (4.6).

$$\tau_L \frac{\mathrm{d}E_p}{\mathrm{d}t} = E_{cmd} - E_p \tag{4.6}$$

The lens absorbs a fixed fraction of every pulse and relaxes when the shutter is closed, as shown in Equation (4.7).

$$\frac{\mathrm{d}T_L}{\mathrm{d}t} = \begin{cases} (1 - T_L)/\tau_{up}, & \text{exposing} \\ -T_L/\tau_{dn}, & \text{otherwise} \end{cases} \tag{4.7}$$

where $T_L \in [0, 1]$ is the normalized heating state. It drives a magnification error $k_L T_L$ in parts per million and is measured by nothing on the tool. The scan coordinate inside the field advances at the scan speed while the field is being written, $\dot{\xi} = v$ during exposure and zero otherwise, and the field is complete when ξ reaches L.

4.3.3 Dose and What Arrives on the Wafer

A point in the field receives the pulses that happen to pass over it while the slit crosses, and because the slit is a full field wide the energy density per pulse is $E_p/(Wh_s)$ while the point sees $f_{rep}h_s/v$ of

them. The slit height cancels, and the delivered dose is Equation (4.8).

$$D = \frac{E_p\, g_p\, f_{rep}}{W\, v}, \qquad N = \frac{f_{rep}\, h_s}{v}, \qquad \frac{\sigma_D}{D} = \frac{\sigma_p}{\sqrt{N}} \tag{4.8}$$

where g_p is the laser gain disturbance, N is the pulse count a point receives and σ_p is the pulse-to-pulse energy scatter. With $f_{rep} = 6000\,\text{Hz}$ and $h_s = 6\,\text{mm}$, N is 90 pulses at 400 mm/s and 51.4 at 700 mm/s. The state actually integrated is the dose accumulated over the whole field scan, $\dot{D}_{acc} = E_p g_p f_{rep}/(WL)$ while exposing and zero otherwise, which is equal to (4.8) at the end of the field and is available to the controller while the field is still being written. The scan ends on a controller period, so the slit can travel past L to ξ_{end} by up to one period of travel (it overshoots by 0.24 mm at 480 mm/s), and the integral is scaled back by L/ξ_{end}, because the dose a point receives does not depend on when the scan stops. Scatter is applied once when the field completes, so the dose error credited to field i is $\Delta D_i/D = D_{acc,i}(L/\xi_{end})(1+\eta_i)/D_{SP} - 1$ with η_i drawn from a normal distribution of standard deviation $\sigma_p/\sqrt{N}$.

The defocus at the slit is the difference between where the stage is and where the recipe wants it, $F = z - (z_w(x,y) + z_{off})$, and the quantity the process responds to combines it with the two terms a stage cannot remove, as shown in Equation (4.9).

$$F_{tot}^2 = F^2 + \frac{1}{3}\left(\frac{W}{2}\,\varepsilon_R\right)^2 + (0.35\,\kappa)^2 \tag{4.9}$$

where $\varepsilon_R = R_x - \theta$ is the across-slit tilt error in radians and $W/2$ is the half field width expressed in µm so that the term carries the units of F. The second term is the mean square of a linear defocus ramp across the field whose value at the edge is $(W/2)\varepsilon_R$, which is what a stage whose slit center is in perfect focus still writes when it has not picked up the surface slope. Equation (4.9) is accumulated only while the shutter is open, and the field result is $F_{rms,i} = \sqrt{\langle F_{tot}^2 \rangle}$ over that interval. Linewidth follows the standard exposure-latitude and Bossung correlation [5, 6], as shown in Equation (4.10).

$$CD_i = CD_0 + s_E \ln\left(1 + \frac{\Delta D_i}{D}\right) - c_F F_{rms,i}^2 \tag{4.10}$$

where s_E is the exposure latitude in nm per natural log of dose and c_F is the Bossung curvature in nm/µm^2. A field is counted as failed when $F_{rms,i}$ exceeds the depth of focus, or the overlay of (4.13) exceeds its specification, or $|CD_i - CD_0|$ exceeds the linewidth tolerance.

4.3.4 Overlay as a Fit to a Sample

The placement error of the printed layer relative to the one underneath it is four correctable terms and one that no linear grid model can represent, as shown in Equation (4.11).

$$\begin{aligned} dx &= T_x + MX - \rho Y + A\sin\frac{2\pi X}{\lambda_{nl}} \\ dy &= T_y + MY + \rho X + A\sin\frac{2\pi Y}{\lambda_{nl}} \end{aligned} \tag{4.11}$$

where X and Y are the field center coordinates in mm, dx and dy are in nm, $M = m + k_L T_L$ is the magnification error in ppm including the lens heating contribution, ρ is the grid rotation in µrad, so that each product with a coordinate in mm is directly in nm, A is the amplitude of the non-correctable distortion and $\lambda_{nl} = 90\,\text{mm}$ is its period. Alignment marks are measured on n_m of the hundred fields, chosen at a stride that spreads them over the wafer rather than clustering them at one radius, and each measurement carries a noise of standard deviation σ_a. Three coefficients per axis are recovered by ordinary least squares through the normal equations, as shown in Equation (4.12).

$$\hat{c}_x = (A_m^\top A_m)^{-1} A_m^\top b_x, \qquad A_m = \begin{bmatrix} 1 & X_k & Y_k \end{bmatrix} \tag{4.12}$$

where A_m carries one row per measured mark, b_x is the vector of measured x displacements at those marks, and the y axis is fitted on $\begin{bmatrix} 1 & Y_k & X_k \end{bmatrix}$ so that the same three-column solver serves both. The fit is performed once, after the marks are measured and before the first field is exposed, and is then applied open loop to all one hundred fields. The residual left on field i is Equation (4.13).

$$e_x = dx - (\hat{c}_{x0} + \hat{c}_{x1} X + \hat{c}_{x2} Y + u_L X) + \sigma_{sync} n \tag{4.13}$$

where u_L is the optional lens heating feedforward of Section 4.5 and σ_{sync} is the per-field stage and reticle synchronization error. The wafer figure of merit is $3\sqrt{\langle e_x^2 + e_y^2 \rangle / 2}$, pooled over both axes, because that is the form an overlay specification is written in.

4.3.5 Sequence and Throughput

The tool runs a state machine of five phases. A wafer exchange takes t_{load}, the alignment phase visits the n_m marks in turn at t_{mark} each, and then one hundred cycles of a step-and-settle of t_{step} followed by an exposure of L/v are executed before the wafer is unloaded. Throughput follows in Equation (4.14).

$$\text{wph} = \frac{3600}{n_m t_{mark} + 100\,(t_{step} + L/v) + t_{load}} \tag{4.14}$$

At the design point Equation (4.14) gives $3600/(6.0 + 16.5 + 0.5) = 156.5$ wph with the reduced 0.5 s exchange the running scenarios use, against a measured 155.8 wph, and 141.2 wph with the default 3.0 s exchange against a measured 140.6 wph. The difference of under half a percent is one 1 ms controller period per field, spent on the transition from step to exposure. The end of the scan is detected at the first controller execution at which the slit has traveled L, within 1 nm of rounding, so at 400 mm/s each field is exposed for exactly 75 ms. Every term in the denominator is a control decision. The alignment time is a sampling plan, the scan time is a focus and dose decision, and both are traded in Section 4.6.

4.3.6 Parameters, Initialization and Numerics

The constants appear in Table 4.2. Three of them never reach the parameter panel, and those three set dynamics a student has to infer from behavior rather than read off a slider.

Every scenario opens with the stage standing on the surface it is about to expose. The height state is initialized to $z(0) = z_w(0,0)$ rather than to the nominal plane, because a clamped 300 mm wafer sits more than a micrometer away from that plane and a stage started at zero begins twenty depths of focus out. The tilt state is initialized to the slope the servo is about to be asked for, $R_x(0) = \theta(0, d_{prev,R})$ with $d_{prev,R}$ the matched tilt lead of Section 4.5, the pulse energy to the open-loop value for the requested dose, and the lens heating and scan coordinate to zero. A shipped numerical test confirms that a stage commanded to the surface it stands on, with the laser at its commanded energy and the shutter closed, has derivatives below 10^{-9} in every state.

The engine of Section 1.6.1 carries the state vector through fourth-order Runge-Kutta on a 0.5 ms fixed step, which leaves a 60 Hz servo well inside the step. Both controllers execute every 1 ms, and 200 rows per simulated second reach the log. Every random draw traces back to seed 20260916, so a swept parameter is the only thing that differs between two runs.

4.3.7 Assumptions

Detail is concentrated on the servo and the sampling, and taken out of everything else. Each simplification has a place where it stops working.

Scalar imaging. Equation (4.10) is a dose and focus correlation, not a lithography simulation. There is no aerial image, no mask topography and no illumination shape, so nothing here reaches optical proximity correction, and the Bossung response is a single quadratic where a real process window is a two-dimensional region in dose and focus whose shape is most of the engineering [5].

Table 4.2 Constants of the Simulated Scanner. The Hidden Three Never Appear on the Parameter Panel, and the Tuning Block Is Shown at Its Shipped Values.

Symbol	Meaning	Value	Unit	Visibility
W	field width	22	mm	fixed
H	field height	24	mm	fixed
R	wafer radius	150	mm	fixed
h_s	slit height along scan	6	mm	adjustable
f_{rep}	laser repetition rate	6000	Hz	adjustable
σ_p	pulse energy scatter	3.0	%	adjustable
b	clamp bow, center to edge	1.5	µm	adjustable
a	long-wavelength shape	0.25	µm	adjustable
ρ_t	short-wavelength topography	0.05	µm	adjustable
L_m	topography correlation length	7	mm	adjustable
h_c	slit chord for slope and curvature	5.5	mm	fixed
z_{lim}	height stage stroke	60	µm	adjustable
R_{lim}	tilt stage stroke	400	µrad	adjustable
DOF	depth of focus, half-width	60	nm	adjustable
CD_0	linewidth at nominal	62	nm	adjustable
s_E	exposure latitude	−24	nm	adjustable
c_F	Bossung curvature	120	nm/µm^2	adjustable
—	linewidth tolerance	2.0	nm	adjustable
—	overlay specification, 3σ	5.0	nm	adjustable
σ_{LS}	level sensor noise	6	nm	adjustable
σ_a	alignment mark noise	1.0	nm	adjustable
σ_{sync}	stage synchronization	0.60	nm	adjustable
t_{mark}	time per alignment mark	0.30	s	adjustable
t_{step}	step and settle between fields	0.090	s	adjustable
t_{load}	wafer exchange	3.0	s	adjustable
k_L	magnification at full heating	0.15	ppm	adjustable
τ_{up}	lens heating time constant	25	s	adjustable
τ_{dn}	lens cooling time constant	180	s	adjustable
L_1	long shape wavelength	95	mm	hidden
a_{max}	height axis acceleration limit	30	m/s^2	hidden
τ_L	laser energy response	0.004	s	hidden
d_{prev}	preview distance	1.70	mm	tuning
d_{trail}	trailing sensor distance	2.0	mm	tuning
f_z, ζ_z	height servo	60, 0.80	Hz	tuning
f_r, ζ_r	tilt servo	45, 0.80	Hz	tuning
$K_{c,f}$	focus trim gain	0.30	µm/µm	tuning
$\tau_{I,f}$	focus trim reset time	0.800	s	tuning
$K_{c,D}$	dose loop gain	8×10^{-4}	—	tuning
$\tau_{I,D}$	dose reset time	0.006	s	tuning
$\hat{k}_L$	lens model coefficient	0.150	ppm	tuning

One height axis and one tilt axis. A production scanner controls six degrees of freedom on the wafer stage and six more on the reticle stage, synchronized to each other to sub-nanometer moving averages [1], and that synchronization carries a large share of the real overlay budget. Here it appears only as the scalar σ_{sync} of (4.13). The assumption removes the cross-coupling that decides how much bandwidth a real stage can be given.

The stage is a clean second order system. Equation (4.4) has no structural resonance, and on a real stage the first flexible mode is usually what limits the achievable bandwidth [1]. The consequence is that raising f_z in this application is free where in a machine shop it is not.

The surface is synthetic. Equation (4.1) is a smooth pseudo-random field rather than a measured wafer, and it carries no topography from the layers underneath. In production, process-induced topography over printed structures is most of the focus problem, and it is not smooth.

Pulses are not simulated individually. The dose integral is continuous and the scatter of (4.8) is applied once when the field completes, so within-field dose structure caused by the servo transient at scan start is absent. The $1/\sqrt{N}$ statement itself is exact, and Section 4.6 verifies it against the model to within 3.8 % at every speed tested.

Alignment is six parameters fitted once per wafer. Production systems fit twenty to forty parameters, apply a correction per exposure field, and close an overlay loop from measurements on the previous lot. The application has none of those, so it overstates how much of the residual a sampling plan alone can remove.

Lens heating drives magnification only. Equation (4.7) is one first-order state. Real lens heating moves several aberrations with different time constants, and it moves focus as well as magnification [15].

The sequence has no stage overlap. Equation (4.14) charges the alignment time to the wafer being written. A dual-stage tool measures the next wafer while the current one is exposed and hides that time entirely, which would change the marks-against-throughput trade of Section 4.6 substantially.

Two disturbances are stylized. A level sensor that acquires a 60 nm bias in one step is not what a real sensor does, but it isolates the one disturbance the feedforward path is structurally blind to. The 0.5 s wafer exchange of the running scenarios is shorter than any real load sequence, and it brings the second wafer quickly to a reader watching at a quarter of real time.

4.4 Why the Preview Distance Belongs to the Scan Speed

The command in Figure 4.1 is built from a surface height measured d_{prev} ahead of the slit, so relative to the point being exposed the command is issued d_{prev}/v seconds early. The plant that answers it is (4.4), whose transfer function is $\omega_z^2/(s^2+2\zeta_z\omega_z s+\omega_z^2)$. Differentiating its phase at the origin gives the low-frequency group delay in Equation (4.15).

$$\tau_g = -\left.\frac{\mathrm{d}\phi}{\mathrm{d}\omega}\right|_{\omega\to 0} = \frac{2\zeta_z}{\omega_z} \tag{4.15}$$

Setting the lead equal to the lag gives the matched preview distance of Equation (4.16), which is the design rule the application puts in front of the reader.

$$d_{prev}^{*} = v\,\tau_g = v\,\frac{2\zeta_z}{\omega_z} = \frac{v\,\zeta_z}{\pi f_z} \tag{4.16}$$

At the design point $\tau_g = 2(0.80)/(2\pi\cdot 60) = 4.244$ ms and (4.16) returns 1.70 mm at 400 mm/s and 2.97 mm at 700 mm/s. Two properties of (4.16) matter more than its value. It is proportional to the scan speed, which is the throughput knob, and it is inversely proportional to the bandwidth, which is the stage design. Neither of the two is a control parameter in the mind of the engineer who moves it.

🕮 **Pause.** Try Exercise 4.1 before reading on.

Equation (4.16) is an estimate and not an identity, because it equates a lead at all frequencies to a group delay that is flat only well below ω_z. The topography this servo is chasing arrives at 57.1 Hz against a bandwidth of 60 Hz, which is exactly the region where the phase is not linear in frequency. Table 4.3 measures the difference by sweeping d_{prev} in 0.05 mm steps over a full wafer at each configuration and recording the focus RMS, and Figure 4.2 plots three of those sweeps.

The first block of Table 4.3 carries the central measurement of this chapter. The measured optimum moves from 1.70 mm at 250 mm/s to 3.65 mm at 700 mm/s, a change of 1.95 mm, while the estimate moves from 1.06 to 2.97 mm, a change of 1.91 mm. The slope is 1.02, so (4.16) scales correctly with scan speed. Across the bandwidth block the estimate falls from 3.40 to 0.85 mm while the measured optimum falls from 3.90 to 1.55 mm, a slope of 0.92 in the same sense.

Table 4.3 Swept Optimum Preview Distance Against the Estimate of Equation (4.16). Each Entry Is the Minimum of a 0.05 mm Sweep Over One Complete Wafer of One Hundred Fields at the Shipped Seed, Where the Surface Is Deterministic and the Sweep Is a Controlled Experiment. F_{rms} at d^* Is Read at the 0.05 mm Setting Nearest d^*, the Resolution the Preview Distance Accepts.

v (mm/s)	f_z (Hz)	ζ_z	d^*_{prev} (mm)	d^{opt}_{prev} (mm)	$d^{opt} - d^*$ (mm)	F_{rms} at d^{opt}	F_{rms} at d^*
250	60	0.80	1.06	1.70	0.64	9.80	12.49
400	60	0.80	1.70	2.35	0.65	11.29	13.76
480	60	0.80	2.04	2.70	0.66	12.39	14.64
550	60	0.80	2.33	3.00	0.67	13.42	15.53
700	60	0.80	2.97	3.65	0.68	16.08	18.05
400	30	0.80	3.40	3.90	0.50	14.92	16.01
400	90	0.80	1.13	1.85	0.72	10.63	13.40
400	120	0.80	0.85	1.55	0.70	10.43	13.51
400	60	0.50	1.06	1.80	0.74	10.43	13.89
400	60	1.20	2.55	3.05	0.50	13.29	14.53

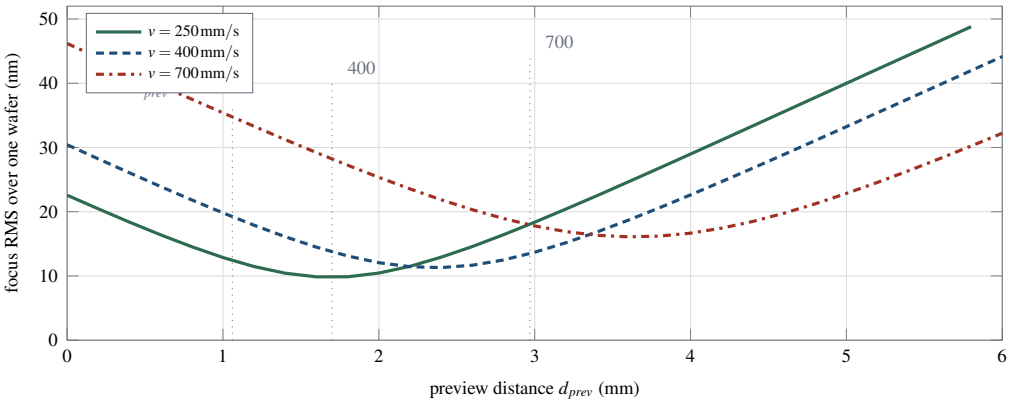


Figure 4.2 Focus RMS over one complete wafer against preview distance at three scan speeds, 250 mm/s solid, 400 mm/s dashed and 700 mm/s dash-dotted, with the estimate of Equation (4.16) marked by the thin gray dotted lines. Every plotted point is one complete wafer of one hundred fields written with the same servo, the same surface and the same seed. The minimum moves out with scan speed as the estimate predicts and sits about 0.65 mm beyond it in all three cases, the lead the trailing-sensor trim takes back, and Table 4.3 locates it on a finer 0.05 mm sweep.

What (4.16) does not give is the offset. The measured optimum exceeds the estimate by between 0.50 mm and 0.74 mm in all ten configurations, and the excess is far more nearly a constant distance than a constant fraction. Stated as a ratio the same data run from 1.15 to 1.83, which makes the ratio the less useful of the two descriptions and makes a quoted percentage correction to (4.16) misleading. The offset does not belong to the servo. It belongs to the trailing-sensor trim of Section 4.5, whose proportional term compares the stage with the surface d_{trail} behind the slit. With the stage on the surface that gap reads d_{trail} times the local surface slope, and the trim subtracts $K_{c,f} d_{trail}$ times the slope from the command, which shortens the effective preview by $K_{c,f} d_{trail} = 0.60$ mm at the shipped gain of 0.30 and distance of 2.0 mm. With the trim gain set to zero the optimum returns to within 0.16 mm of the estimate in all five configurations tried (1.80 mm against 1.70 at 400 mm/s and 3.10 mm against 2.97 at 700 mm/s), and at 400 mm/s the excess follows the trim, 0.40 mm at a gain of 0.15 or a distance of 1.0 mm and 1.15 mm at a gain of 0.60. The gap is a slope only while

d_{trail} is short against the correlation length, which is consistent with the offset falling to 0.40 mm at $L_m = 3$ mm and reaching 0.70 mm at $L_m = 20$ mm with everything else held.

Pause. Try Exercise 4.3 before reading on.

The practical consequence is that (4.16) is a starting point for a search and not an answer. Using the estimate directly at 400 mm/s gives 13.76 nm RMS against 11.29 nm at the swept optimum, and no preview at all, meaning a sensor reading directly under the slit, gives 30.43 nm. The estimate therefore captures 87 % of the available improvement and the search recovers the rest. With the trim gain at zero the estimate itself gives 10.51 nm and the search 10.35 nm, so on a wafer with no sensor bias for the trim to remove it costs 0.9 nm of focus RMS at the optimum.

4.5 Implemented Controllers

Four corrections run at once, and only one of them is a feedback loop closed on the variable it regulates.

4.5.1 Focus: Preview Feedforward With a Trailing Trim

The level sensor reports the surface ahead of the slit with a bias and additive noise, as shown in Equation (4.17).

$$\hat{z}_w = z_w(x,\ y + \sigma_d\, d_{prev}) + \delta + \sigma_{LS}\, n_k \tag{4.17}$$

where $\sigma_d = \pm 1$ is the direction of the current scan, δ is the level sensor bias disturbance and n_k is an independent standard normal draw at each controller execution. The trailing sensor reports the gap behind the slit, $g = z - z_w(x, y - \sigma_d d_{trail}) + \sigma_{LS} n_k$, which is the only true feedback available anywhere in the focus path and is d_{trail}/v seconds old. The command combines the two, as shown in Equation (4.18).

$$z_{cmd} = \mathrm{clamp}\big(\hat{z}_w + z_{off} + u_{trim},\ \pm z_{lim}\big) \tag{4.18}$$

Selecting the trailing architecture replaces $\hat{z}_w$ in (4.18) with zero, which holds the stage at the nominal plane and leaves u_{trim} as the only thing driving it. That switch is the experiment of Section 4.6 and it changes no gain.

The trim is the velocity-form PID of Section 1.6.2, running on the 1 ms controller period Δt_c with the setpoint z_{off} and the process variable g. Each scan computes increments, as shown in Equation (4.19).

$$\begin{aligned} e_k &= a\,(SP_k - PV_k), \qquad a = +1 \\ \Delta u_k^P &= K_c\,(e_k - e_{k-1}), \qquad \Delta u_k^I = \frac{K_c}{\tau_I}\, e_k\, \Delta t_c \\ s_k &= s_{k-1} + \Delta u_k^P, \qquad c_k = s_k + I_{k-1} \\ u_k &= \mathrm{clamp}(c_k,\ u_{min}, u_{max}) \\ I_k &= I_{k-1} + \Delta u_k^I + \frac{\Delta t_c}{T_t}\,(u_k - c_k) \end{aligned} \tag{4.19}$$

Equation (4.19) is Equations (1.12) to (1.15) in this chapter's symbols: s_k is the running sum S_k, c_k is the output before limiting v_k, and the output bias is zero. The action $a = +1$ makes the loop reverse acting in the convention of Section 1.6.2, which is correct here because the gap rises with the command: a stage that is too high reports a gap that is too large and must be commanded down. The integral is seeded at reset with the starting output, zero for the trim and the open-loop pulse energy for the dose loop. Back-calculation corrects only the integral, with the tracking time T_t equal to τ_I. The derivative time τ_D is zero on both loops in this application, so no derivative increment appears. Output limits are ± 20 µm on the trim and no rate limit is applied. The stroke clamp of (4.18) acts after the trim and is not fed back to it, which is immaterial because the command stays within a few micrometers of the nominal plane. Under manual operation I_k is back-solved so that c_k equals the value the operator has dialed in, which makes the transfer bumpless in both directions.

The trim defaults to $K_c = 0.30\,\mu\text{m}/\mu\text{m}$ and $\tau_I = 0.800\,\text{s}$, which is deliberately slower than one field scan of 75 ms. The reason is developed in Section 4.6. Its purpose is to remove a static bias δ in (4.17), which the feedforward path is blind to by construction because the bias is in the sensor the feedforward reads. The trim compares the stage with the surface behind the slit without removing the slope the preview has already followed, so its proportional term also reads the local surface gradient as a focus error, which Section 4.4 measures as a lead of about $K_{c,f} d_{trail}$.

4.5.2 Leveling: Feedforward Only

The tilt command is the measured slope of (4.2) evaluated at the tilt servo's own matched lead, as shown in Equation (4.20).

$$R_{x,cmd} = \text{clamp}\left(\theta\left(x,\; y + \sigma_d \frac{2v\zeta_r}{\omega_r} \right) + \tfrac{1}{4}\sigma_{LS} n_k,\; \pm R_{lim} \right) \tag{4.20}$$

The measurement noise on the slope is taken as a quarter of the level sensor figure applied directly in µrad, which is a shorthand for the slope noise a two-spot sensor produces rather than a quantity derived from the spot geometry. There is no feedback of any kind on tilt. The lead is computed directly from (4.16) applied to the tilt servo rather than exposed as a separate tuning constant, which at the design point is $400 \cdot 2(0.80)/(2\pi \cdot 45) = 2.26\,\text{mm}$. A real tool implements the two leads by delaying one copy of the same sensor signal by a different amount.

4.5.3 Dose: The One Real Loop

The dose loop is the same algorithm of (4.19) with $K_c = 8 \times 10^{-4}$ and $\tau_I = 6\,\text{ms}$, acting on the dose rate rather than on the accumulated dose so that it can correct a field while the field is still being written. The setpoint is $D_{SP} v/L$, the process variable is the delivered dose rate multiplied by the monitor gain g_m, and the output is the pulse energy command bounded to 0.02–1.10 mJ. The upper bound carries the nominal dose to the top of the scan-speed range, because 800 mm/s at 30 mJ/cm^2 needs 0.88 mJ per pulse by (4.8). The reset time is six controller periods, which is aggressive and appropriate because the plant is the single lag of (4.6) with $\tau_L = 4\,\text{ms}$ and the whole opportunity to act lasts 75 ms. Between fields there is nothing to measure, so the loop is placed in tracking on its own output rather than integrating against a measurement it cannot influence, which is the same mechanism that provides bumpless transfer in Section 1.6.2.

4.5.4 Overlay: A Fit Applied Open Loop

The grid correction of (4.12) is computed once per wafer and then applied without further measurement. When the lens heating model is enabled the correction carries one additional term, as shown in Equation (4.21).

$$u_L = \hat{k}_L \left(T_L - T_{L,align} \right) \tag{4.21}$$

where u_L is in ppm and $T_{L,align}$ is the heating state recorded at the moment the marks were measured. Equation (4.21) contains no measurement of the disturbance. It is a model driven by an exposure history the tool already knows, which is the only correct answer available when a disturbance finishes before it could be measured. The float glass line of Volume 5, Chapter 5 meets the same problem in a quality that is fixed before it can be read.

4.6 Results

Table 4.4 lists the eleven scenarios in teaching order, with a short description in place of each name the app shows. A scenario pins the opening mode, the tuning constants, the disturbance settings and whatever is scheduled to happen, and puts its expected observation in front of the reader before the run starts, as Section 1.6.4 describes. The values that follow come off the application's own scorecard and per-wafer export, read at the instant the first wafer completes so that every one of them rests on exactly one hundred fields. That convention matters for the scattered quantities, because the per-wafer

Table 4.4 Scenario Sequence and the Mechanism Each Shows

#	Scenario	What it shows
1	Stage by hand, parked	no flat wafer, manual travel against DOF
2	Servos alone, closed	second-order step response, group delay
3	One hundred fields	the whole sequence and three result maps
4	Trailing sensor only	delay in series with lag, tuning against architecture
5	700 mm/s, preview unchanged	the matched distance belongs to the speed
6	Rougher wafer, 2.2×	spatial frequency against amplitude
7	700 mm/s, 5 % laser	the $1/\sqrt{N}$ dose floor
8	Four alignment marks	fit from a sample, noise into coefficients
9	Lens heating, $k_L = 0.30$ ppm, model pinned off	disturbance finished before measurement
10	Dose monitor +2 % at $t = 8$	a healthy loop making scrap
11	Graded challenge, 480 mm/s	slow servo, short preview, fast trim

counters clear when the next wafer starts, and a run halted at a fixed time estimates a standard deviation from whatever fraction of a wafer has been written by then. Each value is one wafer at the shipped seed unless a table states a twelve-seed mean, and every setting is made before the run starts. The worst-field value is the largest of the hundred per-field values $F_{rms,i}$, not a point value. A setting changed before the run starts re-initializes the case exactly as Reset would (Section 1.6.1), so the dose loop and the laser start at the pulse energy for the speed in force and the application gives one run either way. Figure 4.4 shows the application partway through a wafer of scenario 3.

4.6.1 Manual Operation and the Flat Wafer Assumption

Scenario 1 holds the stage at a fixed height in Manual, and with the tool switched from parked to exposing it writes a wafer under that height. The surface under the slit moves over a range of 1.64 µm across the footprints of the hundred fields, which is 27 times the 60 nm depth of focus. With the stage left at the nominal plane $z = 0$ that the scenario loads, the measured focus RMS is 432.6 nm with a worst field of 955.9 nm. Ninety-nine of the hundred fields fail and the linewidth range across the wafer is 109.3 nm against a 2.0 nm tolerance. Manual mode is not a control strategy here. It is the measurement that establishes the size of the problem, and the manual helper that places the stage on the surface under the slit makes the point that the correct height is correct at exactly one point on one wafer. Parked on the surface at the wafer center, 0.805 µm below the nominal plane, the same wafer measures 684.6 nm, because only 4 % of the surface the slit crosses lies below the center of the clamped wafer.

4.6.2 Architecture Against Tuning

Scenario 4 switches the level sensor out of (4.18) and runs the stage from the trailing sensor alone. At 400 mm/s a trailing distance of 2.0 mm is 5.0 ms of transport delay, and (4.15) adds 4.24 ms of servo lag behind it, on a disturbance whose energy sits at 57.1 Hz. Table 4.5 compares the two architectures on the same wafer, the same servo and the same speed, and Figure 4.3 draws it against the depth of focus.

In the twelve-seed means of Table 4.5, the scenario tuning gives 119.0 nm RMS against a 60 nm depth of focus, a worst field of 251.7 nm and a linewidth range of 7.61 nm on a 2.0 nm specification. Gain helps, and it helps monotonically. Holding the reset time at its shipped 50 ms and raising the trim gain through 0.1, 0.3, 0.6, 1.0, 2.0 and 3.0 gives means of 234.1, 156.4, 119.0, 98.2, 74.2 and 62.8 nm, with no ringing and no reversal anywhere in the range the interface offers. Taking the reset time to its 5 ms minimum as well reaches 41.6 nm and a worst field of 83.8 nm, which is the corner of the permitted range and the best the application allows. The loop does become unstable eventually, and it happens outside the interface, because with the gain bound removed the result on the shipped

Table 4.5 Focus Performance Over One Wafer of One Hundred Fields at 400 mm/s in the Default Configuration. Each Entry Is the Mean of Twelve Seeds, and Worst Is the RMS of the Worst Field. Overlay and Dose Are Inside Specification on Every Row, So the Failure Count Is Attributable to Focus Alone.

Configuration	F_{rms} (nm)	worst (nm)	CD range (nm)	fields failed
Manual, stage at the nominal plane	432.6	955.9	109.29	99.0
Trailing, $K_c = 0.6$, $\tau_I = 50\,\text{ms}$	119.0	251.7	7.61	81.7
Trailing, $K_c = 3.0$, $\tau_I = 50\,\text{ms}$	62.8	134.8	2.26	39.6
Trailing, $K_c = 3.0$, $\tau_I = 5\,\text{ms}$	41.6	83.8	0.96	8.3
Preview at $d^*_{prev} = 1.70\,\text{mm}$	13.7	29.2	0.39	0.0
Preview at $d^{opt}_{prev} = 2.35\,\text{mm}$	11.3	21.0	0.38	0.0
200 Hz servo, noiseless sensor	8.8	19.5	0.43	0.0

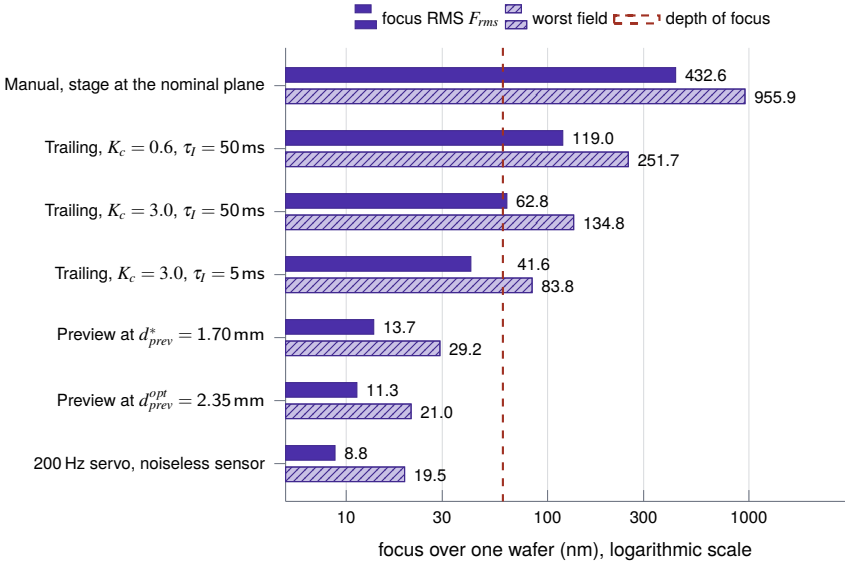


Figure 4.3 Focus RMS (solid bars) and worst field (hatched bars) of the seven configurations of Table 4.5 on a logarithmic scale, with the 60 nm depth of focus as a dashed line. Notice that no tuning of the trailing-sensor loop brings its worst field inside the depth of focus, while both preview rows do, and that the noiseless 200 Hz servo improves on preview only a little.

seed improves further to 47.1 nm at a gain of 8 and then diverges to 1682 nm at 10, while the slider stops at 3.

What gain cannot buy is enough of itself. The corner of the permitted range is 3.0 times worse than preview feedforward on the same servo, the same wafer and the same speed, its worst field of 83.8 nm lies outside the 60 nm depth of focus, and the gain-only result of 62.8 nm is more than four times the preview figure. Shortening the trailing distance helps in the same bounded way, from 119.0 to 109.1 nm at 0.5 mm instead of 2.0 mm, and to 44.6 nm with 18 fields still failing at the gain of 3.0, because the servo lag that remains is larger than the transport delay that was removed. The architecture and not the tuning sets the number that can be reached, and the reason is visible in the arithmetic before any run is made.

Preview feedforward at the estimate of (4.16) gives 13.7 nm RMS and 0.39 nm of linewidth range, and at the swept optimum 11.3 nm and a 21.0 nm worst field. The last row of Table 4.5 establishes how much of that remainder is even addressable. With the height servo at 200 Hz, the tilt servo at

Table 4.6 Throughput Against Focus and Dose. Wafers per Hour Are Computed With the 0.5 s Wafer Exchange the Running Scenarios Use. Dose 3σ Is the Mean of Twelve Seeds at the Default $\sigma_p = 3\,\%$, With the Speed Set Before the Run Starts.

v	wph	N	dose 3σ (%)	F_{rms} at 1.70 mm	F_{rms} at d^{opt}	d^{opt}_{prev} (mm)
250	130	144.0	0.77	9.8	9.8	1.70
400	156	90.0	0.92	13.8	11.3	2.35
480	164	75.0	1.02	17.3	12.4	2.70
550	171	65.5	1.12	20.6	13.4	3.00
700	181	51.4	1.25	28.2	16.1	3.65

150 Hz, the level sensor noise removed and the preview rematched to 1.25 mm, the RMS is 8.8 nm. Most of that residual is the across-slit curvature of (4.3), 7.7 nm on its own, with the tilt term of (4.9) and the slope the trim reads making up the rest, and no bandwidth reaches it. The design configuration is therefore already within 2.5 nm of the floor, which is the answer to a reader who proposes to keep raising f_z.

> ✦ **DID YOU KNOW? Twice the Bandwidth, Almost the Same Focus**
>
> From the simulator: in Table 4.3, with the preview swept to its best value each time, the 60 Hz height servo gives a focus RMS of 11.29 nm and a 120 Hz servo 10.43 nm. Doubling the bandwidth of the stage barely moves the focus, because the curvature across the slit sets a floor that no servo reaches.

4.6.3 Throughput Is a Focus Decision

Scenario 5 raises the scan speed from 400 to 700 mm/s for throughput and changes nothing else. The focus RMS goes from 13.8 to 28.2 nm and the worst field from 29.3 to 63.2 nm, which is outside the depth of focus and fails. Rematching the preview to 3.65 mm returns it to 16.1 nm and 27.6 nm. Table 4.6 carries the whole trade across five speeds.

Going from 400 to 700 mm/s buys 25 wph, an increase of 16 %, and costs 4.8 nm of focus RMS even when the preview is rematched at every step. Leaving the preview alone costs 14.5 nm instead, three times as much, for exactly the same throughput. Nothing on the tool reports that a control parameter moved when the recipe speed changed, which is the point of the scenario.

> **TRY IT IN THE APP 5 · The Preview Distance Belongs to the Scan Speed**
>
> Load the scenario and press Run. The scan speed has been raised to 700 mm/s and the preview left at 1.70 mm. Watch the Focus RMS tile when the wafer unloads: it reads 28.2 nm, against 13.8 nm at 400 mm/s. Set Level sensor preview distance to 3.65 mm and run again: the tile returns to 16.1 nm.

4.6.4 Amplitude Against Spatial Frequency

Scenario 6 multiplies the short-wavelength topography by 2.2 and leaves the clamp bow alone, taking the focus RMS from 13.8 to 21.6 nm. Doubling the same term gives 20.1 nm. Tripling the clamp bow instead, which is a change of 3.0 µm in amplitude against the 0.05 µm added by doubling the topography, gives 23.2 nm with one edge field at 64.5 nm, while the stage travels 4.3 µm during exposure. The servo is not what that wafer costs. With the trim gain at zero the tripled bow measures 13.7 nm against 10.5 nm for the nominal wafer, because the bow arrives at 1.3 Hz and the servo follows it, leaving only its curvature and tilt across the slit. The rest is the trailing sensor reading the bow's slope 2 mm behind the slit as a focus error, largest at the edge where the bow is steepest. The topography arrives at 57.1 Hz and costs about 8 nm whichever way the trim is set (18.7 nm at 2.2× with the trim gain at zero), and raising the bandwidth to 90 Hz with the preview left at 1.70 mm

Table 4.7 Measured Dose 3σ Across One Hundred Fields Against the Statistical Prediction of Equation (4.8). Each Measurement Is the Mean of Twelve Independent Seeds, Each Run Terminated at the Instant the First Wafer Completes, So That Every Estimate Rests on Exactly One Hundred Fields. Speed and σ_p Are Set in Scenario 7 Before Each Run Starts.

σ_p (%)	v (mm/s)	N	measured 3σ (%)	$3\sigma_p/\sqrt{N}$ (%)
3	200	180.0	0.645	0.671
3	400	90.0	0.919	0.949
3	700	51.4	1.250	1.255
5	200	180.0	1.075	1.118
5	400	90.0	1.532	1.581
5	700	51.4	2.083	2.092

recovers only 2.4 nm. With the trim out of the comparison, fifty times the amplitude (3.0 µm against the 0.06 µm that 2.2× adds) costs less than half as much focus. The wafer specification that matters is therefore height change per unit of travel rather than total thickness variation, which is why site flatness is defined over a field and not over a wafer [13].

> **TRY IT IN THE APP** **6 · A Rougher Wafer**
>
> Press Run and read the Focus RMS tile when the wafer unloads: 21.6 nm with the short-wavelength topography at 2.2×, against 13.8 nm for the nominal wafer. Reset, set Chuck clamp bow to 3× and Short-wavelength topography back to 1×, and run again: 23.2 nm, with one edge field at 64.5 nm. Set Focus trim gain to 0 as well, and the bowed wafer gives 13.7 nm.

4.6.5 A Floor No Loop Reaches

Scenario 7 runs at 700 mm/s with a 5 % laser. Table 4.7 compares the measured dose 3σ over one wafer against $3\sigma_p/\sqrt{N}$ from (4.8), averaged over twelve seeds at each condition.

The measured values track the prediction within 3.8 % at every condition in Table 4.7, and the residual is the sampling error of a standard deviation estimated from one hundred fields. The dose loop has no feedforward from the scan speed, so a speed changed during a run reaches the next field with the old pulse energy. Changed from 700 to 400 mm/s after the first wafer, as scenario 7 asks, the first field after the change prints 13 % hot and that wafer reads 4.24 %, which is why the scenario asks for a second wafer after the speed change. Raising the dose loop gain tenfold, to 0.008, moves none of the floor, and the 5 % laser reads 1.49 % at 400 mm/s at either gain on the shipped seed, because by the time a pulse could be corrected it has been delivered to the point and the point has left the slit. The gain range ends at 0.0095, below the clipped oscillation that appears at 0.01 on the 400 mm/s case. The loop is doing its real work elsewhere. When the laser gain drops to 0.93, the loop sees the change on the energy monitor within one field and raises the command, and only the first field written after the drop receives a small part of the change while the loop catches up. Recognizing that a loop has reached a statistical floor is the lesson, and the ways out of it are repetition rate, slit height, laser quality and scan speed rather than tuning [14].

Pause. Try Exercise 4.2 before reading on.

The companion failure is scenario 10. The dose monitor begins reading 2 % high at $t = 8$ s. The loop lowers the pulse energy by the same 2 % and reports the dose exactly on setpoint. Running the whole wafer with the monitor gain at 1.02 moves the mean linewidth from 61.976 to 62.450 nm, a shift of 0.474 nm on every field of every wafer at every seed tried, produced by a fast and correctly tuned loop doing exactly what it was told. In the scenario as loaded the drift arrives as the stage steps to field 10, and the first wafer's mean reads 62.407 nm with no field failed, so nothing on the tool flags it.

Table 4.8 Overlay Residual and Throughput Against the Alignment Sampling Plan. Each Residual Is the Mean of Twelve Seeds, With the Observed Range in Parentheses. The Specification Is 5.0 nm.

n_m	overlay 3σ (nm)	alignment (s)	wph
4	8.84 (4.30–21.76)	1.2	197
8	4.06 (3.31–4.48)	2.4	185
12	3.65 (3.30–4.34)	3.6	174
20	3.22 (2.95–3.45)	6.0	156
32	3.21 (3.05–3.40)	9.6	135
40	3.16 (2.97–3.36)	12.0	124

4.6.6 A Trim Faster Than One Product Unit

The trim reads the physical gap behind the slit, surface slope included, and the preview is a separate feedforward path, so the trim must stay slow enough to correct offsets without chasing the topography. It exists to remove the sensor bias δ of (4.17), which the feedforward path cannot see. Firing a 60 nm level sensor drift at the start of the run with the trim gain at zero gives a focus RMS of 60.8 nm, because the whole wafer is now exposed one full depth of focus away from best focus. Restoring the default gain of 0.30 returns it to 13.6 nm. That is the entire justification for the loop. The coagulant dose of Volume 3, Chapter 6 divides the work the same way, with a fast feedforward from the raw water and the late measurement demoted to a slow trim.

Its reset time then decides whether it helps or hurts, because it is the only element of the focus path with a memory longer than one sample. A field scan at 400 mm/s lasts 75 ms. Setting $\tau_I = 30$ ms makes the trim fast enough to track the surface within a field, and because the trailing sensor sits 2.0 mm behind the slit, driving its gap to the offset inside a field drives the slit itself off best focus by the local surface gradient multiplied by that distance. The measured cost at the design point is a focus RMS of 26.3 nm and a worst field of 51.2 nm, against 13.8 nm and 29.3 nm at the default 800 ms. The same effect decides the graded challenge. With the servo and the preview already repaired, the challenge configuration gives 24.2 nm RMS and a 47.8 nm worst field at $\tau_I = 30$ ms and 11.1 nm and 22.4 nm at 800 ms. Same loop, same gain, a factor of 2.2 in the result, and the rule that produced it is that a feedback trim on a batch process must average over several units rather than act inside one.

Taken from its given state, the challenge scenario runs at 480 mm/s on a 28 Hz servo at 0.45 damping, with the tilt servo at 30 Hz and 0.60, 0.50 mm of preview, a 30 ms trim and 2 nm of level sensor noise. Its first wafer measures 44.0 nm RMS with a worst field of 80.8 nm and 13 fields failed. The challenge is graded over a 45 s run, two wafers at this speed, on the defocus at the center of the slit as logged at 200 Hz, and every sample in the last third of the run (from $t = 29.7$ s, inside the second wafer) must lie within ± 60 nm. The given tuning reaches 131.7 nm and fails. The suggested configuration of 110 Hz at 0.85 damping, the tilt servo at 85 Hz and 0.85, 1.6 mm of preview and an 800 ms trim measures 11.1 nm and 22.4 nm on the first wafer and a worst graded sample of 34.1 nm, inside the band with 25.9 nm to spare. Repairing the servo alone gives 27.9 nm RMS and a worst sample of 81.5 nm, and slowing the trim alone gives 38.5 nm and 116.8 nm, so neither fault can be cleared on its own. The suggested tuning carried through a Reset gives the same run.

4.6.7 Sampling a Spatial Field

Scenario 8 fits the six parameters of (4.12) from four marks. Table 4.8 reports the overlay residual and the throughput against the number of marks in the default configuration, where the lens heating model is on and the sampling effect is therefore the only thing moving. Each entry is the mean over twelve seeds, because a single realization of the mark noise is exactly what a sparse fit amplifies.

Four marks give 8.84 nm against a 5.0 nm specification, and the spread across seeds runs from 4.30 to 21.76 nm. Three parameters fitted from four points leave one degree of freedom, so the mark noise lands almost entirely in the coefficients, and because two of the three coefficients multiply

position the resulting error grows with radius. Eight marks remove most of it at 4.06 nm, and the curve then flattens hard. From twelve marks on, the means of 3.65, 3.22, 3.21 and 3.16 nm differ by less than the spread across seeds, while throughput falls from 174 to 124 wph. The flattening is the non-correctable term A of (4.11) together with the synchronization error σ_{sync}, and no sampling plan reaches below it. Setting A to zero takes twenty marks from 3.22 to 2.06 nm, and setting σ_{sync} to zero as well takes it to 1.04 nm, which leaves only the mark noise and decomposes the floor directly. The optimum sampling plan therefore depends on how much of the real distortion is linear, and the application exposes A as a slider so that the answer can be computed rather than assumed.

4.6.8 A Disturbance That Finishes Before It Is Measured

Scenario 9 lets the lens absorb a fraction of every pulse with $k_L = 0.30$ ppm and $\tau_{up} = 18$ s, so the magnification drifts through the twenty seconds in which the hundred fields are written. The marks were measured before the first field and are not measured again, so no feedback is available at all, and a loop closed on the next wafer would be a wafer too late. The overlay residual is 12.72 nm with the model off and 6.98 nm with the model of (4.21) on, averaged over twelve seeds, and matching the model coefficient $\hat{k}_L$ to the true 0.30 ppm returns 3.22 nm, the residual of a wafer whose lens drift is fully corrected. Nothing measured the disturbance in any case. Lens heating is a first-order rise driven by an exposure history the tool knows exactly, so predicting it is easier than measuring it [15], and the quality of the correction becomes the quality of the model rather than the quality of any tuning.

The same mechanism decides the nominal result, which is why the model is on in the default configuration and scenario 9 is the one place that deliberately pins it off. At twenty marks with the default $k_L = 0.15$ ppm the twelve-seed mean overlay residual is 3.22 nm against the 5.0 nm specification with no field failing. Switching the model off on the same wafers takes the mean to 5.68 nm and fails seven of the hundred fields on overlay alone at the shipped seed (1 to 10 across twelve seeds), while the focus and dose results do not change by a measurable amount. A correction with no measurement anywhere in it is worth 2.46 nm of overlay on a specification of five.

4.7 Technology in Practice

The corrections of Section 4.5 run on two height sensors, an energy monitor, a stage, a laser and the scanner's own controller, drawn together in Figure 4.5.

What Measures It

The feedforward reads the level sensor of Equation (4.17), which the model reduces to a reading of the surface d_{prev} ahead of the slit with a noise of 6 nm and a bias δ.

> **TECHNOLOGY Sensor: An Oblique Beam on the Wafer**
>
> A level sensor sends a light beam onto the wafer at a grazing angle, of the order of 80° from the normal, and images the spot it makes onto a detector of two cells. When the surface lies in the image plane the returned spot falls symmetrically and the difference of the two cells is zero. A surface too high or too low moves the spot, and the sign and size of the difference give the direction and size of the error [18]. The wafer itself can fool the sensor. A wafer under several layers of resist reads at a different height from one under a single layer, though both sit at the same height, an error called process dependency [19]. On a focus trend it is a steady offset, invisible to a feedforward that reads the same sensor.

Process dependency is the kind of error the bias δ stands for, and the stylized 60 nm step of the level sensor drift is the case the trailing-sensor trim exists to remove. The dose loop reads an energy monitor whose gain g_m is the disturbance of scenario 10.

What Moves It

The height and tilt axes are second-order servos of 60 Hz and 45 Hz at 0.80 damping, with a stroke of 60 µm, an acceleration limit of 30 m/s^2 and no structural resonance (Equation (4.4)).



Figure 4.4 Scenario 3 partway through a wafer at $t = 14.1\,\text{s}$, with field 46 of 100 being exposed at 400 mm/s and the preview distance matched at 1.70 mm, cropped to the Scan view and the first two trend strips. The Scan view is a section through the slit at a vertical scale of 10 nm per pixel, with the level sensor spot ahead of the slit (1), the trailing spot behind it (2) and the ±60 nm depth of focus drawn as a band (3). The stage height command (5) sits at −0.177 µm against a stage at −0.185 µm while the defocus at the slit reads +8.6 nm (4), and the scorecard chips above the view report a focus RMS of 14.2 nm, an overlay 3σ of 3.28 nm and a dose 3σ of 0.82 % over the forty-five fields written so far, with no field failed. Throughput reads zero because no wafer has completed yet.

TECHNOLOGY **Final Element: A Levitated Wafer Stage**

The stage is driven by linear and planar motors on the Lorentz principle. Their permanent magnets give a field of about 0.7 T at the moving coils, whose current density is pushed to several tens of amperes per square millimeter by water cooling [7]. Moving without contact, by magnetic levitation, avoids friction and improves reproducibility, and in vacuum it avoids the wear and lubricants that would pollute the tool. Throughput rises as the moving mass falls, but a lighter stage bends, and its flexible modes then reach into the frequencies the loop must control [20]. Each degree of freedom is held by its own PID loop, with notch filters on the resonances that endanger stability [7]. An acceleration feedforward, the estimated mass times the reference acceleration, leaves the feedback little to correct [20].

The second final element is the laser, which follows its energy command through a lag of 4 ms within 0.02 to 1.10 mJ per pulse (Equation (4.6)). Both PI loops use $a = +1$ and are reverse acting in the convention of Section 1.6.2, because each process variable rises with its output.

Where the Controller Runs

The focus and dose loops run on the scanner's own real-time controller. Control loops in operational technology run with cycle times from milliseconds to minutes, many on real-time operating systems,

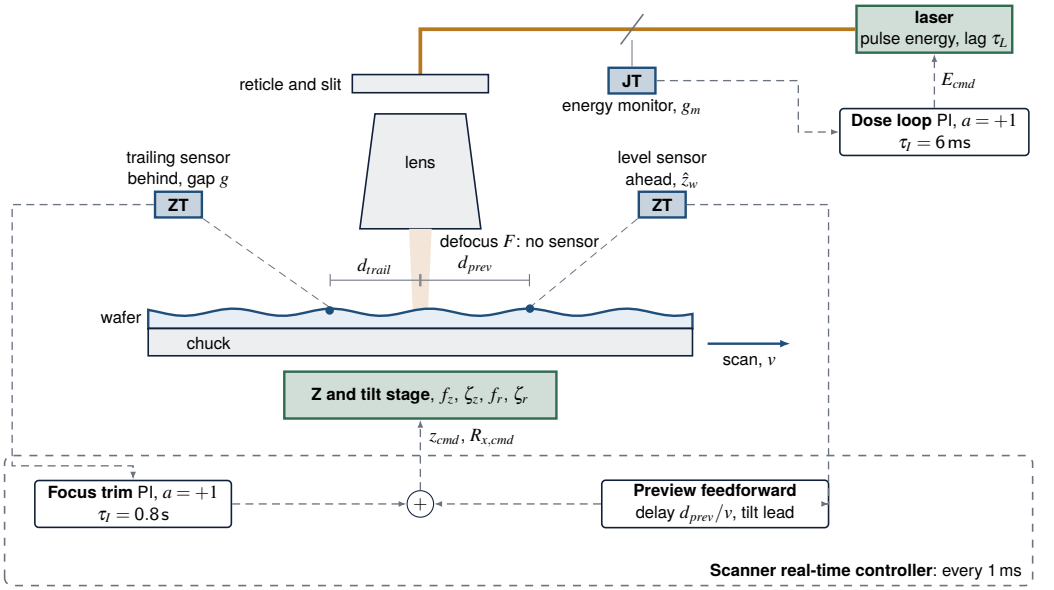


Figure 4.5 The focus and dose loops of the simulated scanner drawn with their hardware. The level sensor reads the surface d_{prev} ahead of the slit, and its reading, delayed by the scan, becomes the height command z_{cmd}. The trailing sensor reads the gap d_{trail} behind the slit and feeds the slow focus trim, a PI loop with $a = +1$. The energy monitor feeds the fast dose loop, which sets the pulse energy of the laser. Both loops run every 1 ms. Notice that no instrument looks at the slit itself, where the defocus F is decided.

and what counts as real time depends on the application and has to be stated for each [21]. The stage that executes the height command is held by one PID loop with notch filters for each degree of freedom [7]. The app executes both of its loops every 1 ms against a physics step of 0.5 ms, and a field scan at 400 mm/s lasts 75 ms. In the controller the structure is three blocks: a delay on the level-sensor signal that turns the preview distance into a lead time, a slow PI block on the trailing sensor, and a fast PI block on the energy monitor. Corrections that need measurements of the printed wafer run outside the scanner, lot by lot, in the run-to-run loops of Section 1.4.3 and Chapter 6.

> **IN PRACTICE** **Commissioning a Stage Loop**
>
> A stage loop is tuned on measurements. The engineer measures the frequency response of the stage, decouples it so that each degree of freedom can be handled alone, and shapes a PID controller with notch filters by hand on that measured response. The feedforward, mass times the reference acceleration, is then tuned by removing the correlation between the measured error and the acceleration profile, and learning control may refine it afterward [20]. The procedure holds while the stage moves as a rigid body over the frequencies that matter. This chapter's height servo, a bandwidth and a damping ratio, is what such a procedure hands to the focus layer above it.

What Is Modernizing It

Three developments bear on this tool. The first is overactuation: more actuators and sensors than the stage has rigid-body degrees of freedom, used to stiffen and damp its flexible modes. It has entered industrial wafer stages over the last five to ten years [7], and on a prototype stage one added actuator and sensor pair raised the bandwidth by 35 % [20]. The second is the reuse of identified models in real time, as digital twins that detect and isolate faults and support predictive maintenance [7]. The third attacks the surface itself. The unflatness of the wafer table is one of the main contributors to

focus variation, and a table that deforms the wafer edge with over 1600 embedded piezoelectric actuators improved from over 200 nm to under 15 nm in simulation [7].

4.8 Limitations

The application is not an imaging simulator and cannot be used to choose a process. Equation (4.10) is a two-term correlation where a real process window is a surface in dose and focus whose shape decides the exposure latitude and the usable depth of focus [5, 6]. The wafer surface is synthetic and carries no process-induced topography, which in production is most of the focus budget. One height axis and one tilt axis stand in for twelve controlled degrees of freedom across two stages, so the synchronization that dominates a real overlay budget appears only as a scalar noise term [1]. The stage has no structural resonance, so the bandwidth in this model is free where in a machine it is bought with mass, stiffness and a great deal of identification work [1]. Alignment is six parameters fitted once per wafer with no per-field correction and no lot-to-lot loop of the kind Chapter 6 closes on linewidth, which makes the sampling trade of Table 4.8 cleaner than the real one. Pulses are not simulated individually, so within-field dose structure from the scan-start transient is absent even though the $1/\sqrt{N}$ floor is reproduced to 3.8 %. Throughput charges the alignment time to the wafer being written, and a dual-stage tool hides it. Exact repeatability is itself a fiction of the simulator, because a fixed step and a seeded generator hand the same trajectory to every machine, which is not what a production tool does. What survives the simplification is the structure of the problem rather than the machine. The matched preview relation, the $1/\sqrt{N}$ dose floor, the spatial frequency argument and the fit-from-a-sample trade are statements about any process that transports material past a fixed working point, and they are what transfers [2, 3]. Section 4.3.7 says of each assumption in turn where it stops holding.

> **OPEN CHALLENGE** **The Error Where the Light Falls**
>
> On a wafer stage the accuracy that matters is at the spot of exposure, while the position sensors measure the edge of the stage. A lighter stage would raise throughput, but it would bend within the control bandwidth, and the error at the exposure spot would no longer follow from the error at the sensors. Inferring that unmeasured error, from more sensors, more actuators and a model, is an open problem, and the lack of control methods for it is the main reason such lightweight stages are not yet feasible [20]. Progress would be a stage whose error at the exposure spot is estimated and held as well as its sensor reading is today.

4.9 Discussion

The coupling this application makes visible has a commercial name. A scanner is the most expensive tool in a fab and its output is counted in wafers per hour, so the scan speed is under continuous pressure from the side of the organization that measures throughput. Table 4.6 shows what that pressure buys and what it spends. Between 400 and 700 mm/s the tool gains 25 wph and spends 4.8 nm of focus RMS with the preview rematched and 14.5 nm without, while the dose 3σ of Table 4.7 rises from 0.92 to 1.25 % at a 3 % laser and from 1.53 to 2.08 % at a 5 % one. None of those three costs appears on a throughput report, and two of the three are recoverable only by changing a control parameter that the throughput decision never mentioned. The alignment sampling plan of Table 4.8 is the same trade written the other way around, where 10.8 s of extra measurement buys 5.7 nm of overlay and costs 73 wph.

> **OPEN CHALLENGE** **Wafers That Will Not Lie Flat**
>
> Deposition, thermal processing and polishing leave stress that warps a wafer, and memory chips of 150 to 300 exposed layers already bring up to 0.5 mm of warpage. How a warped wafer settles on its table depends on friction that changes with wear and contamination, so the result does not repeat and leaves overlay errors [7]. Progress would be a clamped surface that repeats from wafer to wafer.

The model of Section 4.3 supports control methods beyond the loops implemented here. The preview command of (4.18) is the simplest member of a family in which the whole measured surface ahead of the slit, rather than one point of it, is the input to an optimization over the stage dynamics and its acceleration limit [4]. A model predictive formulation on the same eight states, of the kind Section 1.6.3 describes, would enforce the stroke and acceleration constraints of Equation (4.4) explicitly and tell the reader directly what the constraint costs. The wafer is written a hundred times per lot with the same stage and a similar surface, which is the setting iterative learning was built for, as in the powder bed of Volume 8, Chapter 10, and the reason feedforward on these machines is identified from repeated trials rather than tuned [1]. The alignment fit of (4.12) generalizes to a regularized estimator in which the mark noise and the expected distortion enter as covariances, which is what turns Table 4.8 from a sweep into a design.

4.10 Conclusion

This chapter studied a step-and-scan exposure sequence as a laboratory for control at a point that cannot be instrumented. The simulator integrates eight states and writes one hundred fields on a 300 mm wafer in about twenty seconds. Focus is commanded from a level sensor reading d_{prev} ahead of the slit, and the matched distance $v\,2\zeta_z/\omega_z$ is 1.70 mm at 400 mm/s on a 60 Hz servo at 0.8 damping. A wafer exposed from a trailing sensor instead measures 119.0 nm of focus RMS and 7.61 nm of linewidth range, and raising the trim gain improves that monotonically to 62.8 nm at the top of its slider and 41.6 nm at the corner of the permitted range, while the same servo run from the sensor ahead measures 13.7 nm and 0.39 nm against a floor of 8.8 nm that the across-slit curvature sets. A sweep places the true optimum between 0.50 and 0.74 mm beyond the group-delay estimate in all ten configurations tested, while the estimate tracks the scaling with speed at a slope of 1.02 and with bandwidth at 0.92. The excess is the lead the trailing-sensor trim takes back, $K_{c,f}d_{trail} = 0.60$ mm, and with the trim gain at zero the optimum falls to within 0.16 mm of the estimate. Raising the speed to 700 mm/s without rematching takes the focus RMS to 28.2 nm and the dose 3σ from 0.92 to 1.25 % because a point receives 51.4 pulses instead of 90. Slowing the trailing-sensor trim from a reset time of 30 to 800 ms takes the design point from 26.3 to 13.8 nm, because a feedback trim tuned faster than one product unit converts a surface gradient into a focus error. A fuller model would add a structural resonance to the stage, a per-field overlay correction, process-induced topography from a measured wafer, and a second stage so that alignment time can be overlapped. The scanner shows that feedback is a choice wherever the disturbance can be seen before it arrives.

KEY POINTS

- No sensor can sit where the light falls, so the stage is commanded from a level sensor that reads the surface ahead of the slit.
- The matched preview distance, $v\,2\zeta_z/\omega_z$, is 1.70 mm at 400 mm/s and moves whenever the scan speed is raised for throughput.
- Fed from a trailing sensor, the same servo measures 119.0 nm of focus RMS and 41.6 nm at the corner of the permitted range, against 13.7 nm with preview.
- The swept optimum sits 0.50 to 0.74 mm beyond the estimate, the lead the trailing-sensor trim takes back.
- A feedback trim tuned faster than one product unit converts a surface gradient into a focus error.

Where else this applies. Preview from a sensor ahead of the working point in Volume 7, Chapter 11 and Volume 20, Chapter 12, and a late measurement demoted to a slow trim in Volume 3, Chapter 6.

References

[1] H. Butler, “Position control in lithographic equipment [applications of control],” *IEEE Control Systems*, vol. 31, no. 5, pp. 28–47, 2011.

[2] M. Tomizuka and D. H. Fung, "Design of digital feedforward/preview controllers for processes with predetermined feedback controllers," *Journal of Dynamic Systems, Measurement, and Control*, vol. 102, no. 4, pp. 218–225, 1980.

[3] A. Hać, "Optimal linear preview control of active vehicle suspension," *Vehicle System Dynamics*, vol. 21, no. 1, pp. 167–195, 1992.

[4] M. Tomizuka and D. E. Rosenthal, "On the optimal digital state vector feedback controller with integral and preview actions," *Journal of Dynamic Systems, Measurement, and Control*, vol. 101, no. 2, pp. 172–178, 1979.

[5] C. Mack, *Fundamental Principles of Optical Lithography*. Chichester, UK: John Wiley & Sons, 2007.

[6] J. W. Bossung, "Projection printing characterization," in *Developments in Semiconductor Microlithography II*, ser. Proceedings of SPIE, vol. 0100. SPIE, 1977, pp. 80–85.

[7] M. Steinbuch, T. Oomen, and H. Vermeulen, "Motion control, mechatronics design, and Moore's law," *IEEJ Journal of Industry Applications*, vol. 11, no. 2, pp. 245–255, 2022.

[8] J. H. Bruning, "Optical lithography: 40 years and holding," in *Optical Microlithography XX*, ser. Proceedings of SPIE, vol. 6520. SPIE, 2007, p. 652004.

[9] K. Suwa, K. Ushida, and B. J. Lin, "The optical stepper with a high numerical aperture i-line lens and a field-by-field leveling system," in *Optical/Laser Microlithography*, ser. Proceedings of SPIE, vol. 0922. SPIE, 1988, pp. 270–276.

[10] M. A. van den Brink, B. A. Katz, and S. Wittekoek, "New 0.54 aperture i-line wafer stepper with field-by-field leveling combined with global alignment," in *Optical/Laser Microlithography IV*, ser. Proceedings of SPIE, vol. 1463. SPIE, 1991, pp. 709–724.

[11] K. Suzuki, S. Wakamoto, and K. Nishi, "KrF step-and-scan exposure system using higher-NA projection lens," in *Optical Microlithography IX*, ser. Proceedings of SPIE, vol. 2726. SPIE, 1996, pp. 767–779.

[12] P. K. Govil, J. G. Tsacoyeanes, R. P. Eron, and D. Walters, "Contributors to focal plane nonuniformity and their impact on linewidth control in DUV step-and-scan system," in *Optical Microlithography XI*, ser. Proceedings of SPIE, vol. 3334. SPIE, 1998, p. 92.

[13] L. Liauw, A. Muray, and M. Chen, "Impact of wafer flatness on submicron optical lithography," in *Optical Microlithography VI*, ser. Proceedings of SPIE, vol. 0772. SPIE, 1987, p. 232.

[14] D. H. Tracy, F. Y. Wu, and B. J. Lin, "Exposure dose control techniques for excimer laser lithography," in *Optical/Laser Microlithography*, ser. Proceedings of SPIE, vol. 0922. SPIE, 1988, pp. 437–443.

[15] Y. Cui, "Fine-tune lens-heating-induced focus drift with different process and illumination settings," in *Optical Microlithography XIV*, ser. Proceedings of SPIE, vol. 4346. SPIE, 2001, p. 1369.

[16] C. Wagner and N. Harned, "Lithography gets extreme," *Nature Photonics*, vol. 4, no. 1, pp. 24–26, 2010.

[17] M. Uehara, T. Sudo, and F. Kanatani, "Horizontal position detecting device," U.S. Patent 4,558,949, filed 17 September 1982, 1985.

[18] J. E. van der Werf, M. A. van den Brink, H. F. D. Linders, and J. M. M. Beltman, "Imaging apparatus having a focus-error and/or tilt detection device," U.S. Patent 5,191,200, filed 16 December 1991, 1993.

[19] P. A. A. Teunissen, R. M. G. J. Queens, P. A. M. Dekkers-Rog, and A. C. F. A. van Well, "Level sensor for lithographic apparatus," U.S. Patent 7,265,364, filed 10 June 2004, 2007.

[20] T. Oomen, "Advanced motion control for precision mechatronics: Control, identification, and learning of complex systems," *IEEJ Journal of Industry Applications*, vol. 7, no. 2, pp. 127–140, 2018.

[21] K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, A. Hahn, S. Saravia, A. Sherule, and M. Thompson, "Guide to operational technology (OT) security," National Institute of Standards and Technology, Gaithersburg, MD, Special Publication 800-82r3, 2023.

4.11 Exercises

4.1 ⌕ Concept check · 2 min

Scenario 5 raises the scan speed from 400 to 700 mm/s and leaves the level sensor preview at 1.70 mm. The focus RMS rises from 13.8 to 28.2 nm, although the servo and the wafer are unchanged. Which explanation is right?

(a) The level sensor noise grows with scan speed, so the command carries more noise at the higher speed

(b) The stage lags its command by a fixed time, which now spans more wafer travel than the preview supplies

(c) The pulse count falls with speed, and the larger dose scatter is converted into a focus error

(d) The servo bandwidth falls when the stage moves faster, so it no longer follows the clamp bow

4.2 Concept check · 2 min

Scenario 7 runs a laser with 5 % pulse-to-pulse energy scatter. At 400 mm/s the Dose 3σ tile reads 1.49 % for a wafer with the dose loop gain at its default of 0.0008 and again with the gain at 0.008. Why does ten times the gain change nothing?

(a) The pulse energy sits on its upper limit, so the extra gain has no room to act

(b) Between fields the loop is held in tracking, so it never sees the pulse scatter

(c) Each point averages the pulses it receives, and each pulse lands before it can be corrected

(d) The energy monitor reads high, and no gain can remove a bias in the measurement

4.3 Concept check · 2 min

At 400 mm/s on the 60 Hz servo at 0.80 damping, a sweep finds the best preview at 2.35 mm, 0.65 mm beyond the group-delay estimate of 1.70 mm. With the focus trim gain at zero the best preview is 1.80 mm. What produces most of the 0.65 mm?

(a) The servo phase near 57.1 Hz departs from its low-frequency group delay and calls for extra lead

(b) The tilt servo is slower than the height servo, so the height command must lead to cover the tilt lag

(c) The stage filters the level sensor noise, and the filter adds a fixed lag that the preview must cover

(d) The trim reads the surface slope behind the slit and shortens the effective lead by about $K_{c,f} d_{trail}$

4.4 Theory and calculation · 5 min

The height servo of the design point has $f_z = 60\,\text{Hz}$ and $\zeta_z = 0.80$, and the tilt servo has $f_r = 45\,\text{Hz}$ and $\zeta_r = 0.80$. The Level sensor preview distance slider moves in steps of 0.05 mm.

(a) Find the low-frequency group delay of the height servo. *Checkpoint:* 4.244 ms.

(b) Find the matched preview distance at 400 and at 700 mm/s, and the slider setting nearest each. *Checkpoint:* 1.698 mm, set as 1.70 mm, and 2.971 mm, set as 2.95 mm.

(c) Find the lead the app computes for the tilt servo at 400 mm/s. *Checkpoint:* 2.26 mm.

4.5 Theory and calculation · 6 min

The laser fires at $f_{rep} = 6000\,\text{Hz}$ through a slit $h_s = 6\,\text{mm}$ tall along the scan, and the field is $W = 22\,\text{mm}$ wide.

(a) Find the number of pulses a point receives at 400 and at 700 mm/s. *Checkpoint:* 90 and 51.4.

(b) For a pulse-to-pulse scatter $\sigma_p = 5\,\%$, find the dose floor $3\sigma_p/\sqrt{N}$ at both speeds and the ratio of the second to the first. *Checkpoint:* 1.58 % and 2.09 %, ratio 1.32.

(c) Find the pulse energy that delivers 30 mJ/cm^2 at 400 and at 800 mm/s, and compare the second with the 1.10 mJ limit of the laser. *Checkpoint:* 0.44 mJ and 0.88 mJ.

4.6 Theory and calculation · 6 min

Linewidth follows Equation (4.10) with $CD_0 = 62\,\text{nm}$, $s_E = -24\,\text{nm}$ and $c_F = 120\,\text{nm}/\mu\text{m}^2$. A field fails if F_{rms} exceeds the 60 nm depth of focus or $|CD - CD_0|$ exceeds 2.0 nm.

(a) In scenario 10 the dose monitor reads 2 % high and the loop holds the measured dose on setpoint. Find the relative dose error delivered and the linewidth shift it causes. *Checkpoint:* −1.96 % and +0.475 nm.

(b) Find the linewidth loss of a field whose F_{rms} equals the depth of focus, and the F_{rms} at which focus alone would use the whole 2.0 nm tolerance. *Checkpoint:* 0.432 nm and 129 nm.

(c) The stage stays tilted 5 µrad from the surface across a whole field. Find the contribution of the tilt term of Equation (4.9) to F_{rms}. *Checkpoint:* 31.8 nm.

4.7 App challenge · 8 min

Load scenario 4, Feedback only, and why it is not enough, with the default seed, and set the playback speed to 1 s/s. Each run below is one wafer. Run until the Throughput tile first shows a value, at $t = 23.1\,\text{s}$, pause between 23.5 and 29 s, and read the Focus RMS tile within 0.1 nm. Make each change before Run, or press Reset after making it.

(a) Run as loaded: Focus is commanded from The sensor behind, Focus trim gain 0.60 and Focus trim integral time 0.050 s. *Checkpoint:* 119.0 nm.

(b) Set Focus trim gain to 3.00, the top of its slider, and run. Then also set Focus trim integral time to 0.005 s, its minimum, and run again. *Checkpoint:* 62.9 nm, then 41.7 nm.

(c) Set Focus is commanded from to The sensor ahead, the trim gain to 0.30 and the trim integral time to 0.800 s, and run. Find the ratio of the second reading of part (b) to this one, and say what it shows. *Checkpoint:* 13.8 nm, ratio 3.0.

4.8 App challenge · 8 min

Load scenario 5, The preview distance belongs to the scan speed, with the default seed, and set the playback speed to 1 s/s. Each run below is one wafer. Run until the Throughput tile first shows a value, at $t = 19.9\,\text{s}$, pause between 20 and 26 s, and read the Focus RMS tile within 0.1 nm. Make each change before Run, or press Reset after making it.

(a) Run as loaded, at 700 mm/s with the preview at 1.70 mm. Read Focus RMS and Throughput. *Checkpoint:* 28.2 nm and 181 wph.

(b) Set Level sensor preview distance to 2.95 mm, the slider step nearest the matched distance $v\tau_g$, and run. Then set it to 3.65 mm and run again. *Checkpoint:* 18.1 nm, then 16.1 nm.

(c) Set the preview back to 2.95 mm and Focus trim gain to 0, and run. Explain why the matched distance now does as well as 3.65 mm did with the trim on. *Checkpoint:* 16.1 nm.

4.9 App challenge · 9 min

Load the graded challenge, Challenge · 480 mm/s on a slow, soft stage, with the default seed and playback speed 1 s/s. It scans at 480 mm/s on a 28 Hz height servo at 0.45 damping with a 0.50 mm preview and a Focus trim integral time of 0.030 s.

(a) Find the group delay and the matched preview at 480 mm/s for the given servo and for a repaired one at 110 Hz and 0.85 damping. *Checkpoint:* 5.12 ms and 2.46 mm, then 2.46 ms and 1.18 mm.

(b) Before Run, set Z servo bandwidth 110 Hz, Z servo damping 0.85, Tilt servo bandwidth 85 Hz, Tilt servo damping 0.85 (in Advanced) and Level sensor preview distance 1.60 mm. Run, pause between 22 and 28 s and read the Focus RMS tile, which then holds the first wafer. Run on past 45 s and read the challenge checklist. *Checkpoint:* 24.2 nm within 0.1 nm. Not met, worst defocus 69.3 nm.

(c) Set Focus trim integral time to 0.800 s, press Reset and repeat. Pause between 45 and 50 s, export the CSV file, and find the largest magnitude of column `defoc_nm` in the rows from 29.7 to 45 s (the time column is in seconds). *Checkpoint:* 11.1 nm. Met, with 34.1 nm within 0.1 nm.

4.10 Application area · 4 min

Equation (4.14) charges the alignment time to the wafer being written. A dual-stage tool measures the marks of the next wafer while the current wafer is exposed.

(a) With 20 marks at 0.30 s each, a 90 ms step and settle, 400 mm/s and the 0.5 s exchange, find the throughput from Equation (4.14), and again with the alignment term removed as a dual-stage tool would remove it. *Checkpoint:* 156.5 wph and 211.8 wph.

(b) On this tool, 40 marks cost 73 wph against 4 marks (Table 4.8). What would they cost on the dual-stage tool, and what then limits the number of marks? *Checkpoint:* Nothing, while 12.0 s of alignment fits inside the 16.5 s of stepping and scanning.

4.11 Application area · 3 min

A wafer supplier offers two upgrades at the same price: halve the 1.5 µm clamp bow, or halve the 0.05 µm short-wavelength topography of 7 mm correlation length. The scanner runs at 400 mm/s with preview feedforward.

(a) Find the frequency at which each term reaches the servo, taking the bow as one wavelength across the 300 mm wafer, and the product of amplitude and frequency squared for each. *Checkpoint:* 1.33 Hz and 57.1 Hz, then 2.67 and 163 µm Hz^2.

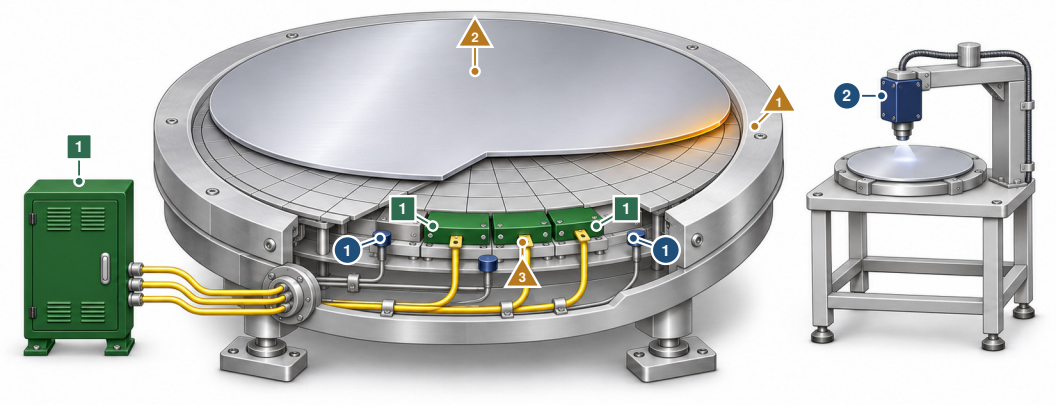
(b) Which upgrade buys focus, and what kind of flatness specification should the purchase order name?

4.12 Review the agent · 5 min

An AI assistant is asked for the preview distance after a recipe change from 400 to 700 mm/s on the design servo, and for a new 120 Hz stage at 400 mm/s. It writes: “The swept optimum at 400 mm/s is 2.35 mm, which is 1.38 times the group-delay estimate of 1.70 mm. The formula therefore underestimates the lead by 38 %. Apply the same factor everywhere: use $2.97 \times 1.38 = 4.10$ mm at 700 mm/s and $0.85 \times 1.38 = 1.17$ mm, set as 1.20 mm, on the 120 Hz stage.” Find the flaw, give better settings, and say how you would check them in the app.

CHAPTER 5

A Hundred-Zone Bake Plate Simulator for Decentralized Control and Spatial Bandwidth



The controller moves 1 a heater under each of the hundred zones, switched from a power cabinet

The controller reads 1 a thermocouple in each zone of the plate

2 a station that measures the printed linewidth on every field

What disturbs it 1 a bowed wafer that touches the plate badly at its edge 2 a cold wafer landing on the hot plate

3 a heater that loses most of its power

Notice that the thermocouples sit in the plate, not in the wafer: only the station at the right sees what the bake did to the printed lines. Illustration, not to scale.

IN PLAIN TERMS A Hot Plate Cut Into a Hundred Squares

The main idea. Light draws a pattern on a thin disk of silicon. Then the disk is baked on a hot plate, so a reaction in its coating can finish. The plate is cut into a hundred heated squares, and each has its own thermometer and controller.

The challenge. Each square shares heat with its neighbors. So a controller set up on one square overreacts when all of them change at once. The plate can also read perfectly while the disk bakes unevenly, because a bent disk touches the plate badly at its edge.

Why it matters. A tenth of a degree changes the width of the printed lines, so the disk itself has to be measured.

THREE WAYS TO READ THIS CHAPTER

Explore	11 min	The first page, the history box in Section 5.2, the try-it and fact boxes of Section 5.6, and the key points
Study	60 min	The whole chapter with its pause points and the twelve exercises
Practice	25 min	The first page, Section 5.6, Section 5.7, Section 5.8, and the open challenge at the end of Section 5.9

The control question A hundred zone loops on one conducting plate can each be well tuned and still hold a temperature profile nobody asked for, while the wafer on top prints a linewidth no thermocouple sees. Which of the plate's disturbances does feedback answer, and which need a model, a timed feedforward or a measurement of the product?

Learning objectives

- Derive the steady gain of the plate against the spatial wavenumber of a power pattern and explain why the zone count does not set the correction the plate can deliver.
- Explain why a gain tuned on one zone overshoots when every zone moves, and compute the factor from the conductances.
- Compute the forward-model power that holds a flat profile at the edge of the plate, and state what it buys that feedback does not.
- Size a wafer-drop pulse on the linewidth rather than on the plate dip, and explain the integral hold.
- Judge a hundred-channel run-to-run correction on its linewidth range and its spatial roughness, with and without a spatial filter.

Prerequisites Sections 1.4.5 (multizone thermal control), 1.4.3 (run-to-run correction), 1.6.2 (velocity form and loop action) and 1.6.4 (scenarios, scorecard and CSV).

The app https://apmonitor.com/apps/peb-bake-zones/

See also Volume 6, Chapter 2 and Volume 6, Chapter 3 for pairing and static decoupling. Volume 8, Chapter 10 and Volume 8, Chapter 9 for corrections learned from late measurements. Volume 3, Chapter 6 for feedforward that leaves feedback only a trim.

5.1 Introduction

One hundred heater zones are cut into one piece of aluminum, each with its own thermocouple, its own power supply and its own loop. Every zone conducts heat to the four around it. That single sentence contains the whole difficulty, because a hundred loops that are individually well tuned can still produce a temperature profile that nobody asked for, and the mechanism is not interaction in the usual pairing sense. The relative gain array of this plant is close to the identity, so the pairing question that a multivariable course asks first is already answered [1]. What is not answered is what happens when the hundred loops move in the same direction at the same time.

The reason they are there at all is chemistry. A chemically amplified resist does not finish reacting when the exposure stops, because the photoacid generated by the light deprotects the polymer during a bake that follows, and the reaction is Arrhenius [2, 3]. What the resist integrates is a thermal dose rather than a temperature, and the linewidth that results is sharply sensitive to the bake, on the order of several nanometers per degree in published measurements [4]. A plate specification written to a tenth of a degree follows directly from that number, and a plate with many independently driven zones follows from the specification.

The class of plant is larger than the application. A diffusive medium carrying many collocated actuators and sensors describes a multizone rapid thermal processing chamber [5, 6] (Chapter 9), a chemical vapor deposition susceptor, an extruder die, an injection mold and a battery electrode dryer. Every one of them has the property that matters here. The steady-state map from a power pattern to a temperature pattern is not one gain but a family of gains indexed by spatial wavenumber, which is the defining feature of a spatially invariant distributed system [7]. A plate with more zones cut into it resolves a finer correction and does not necessarily produce one. However, a reader handed a 100×100 gain matrix and told to design a decentralized controller learns the linear algebra and never meets the consequence, which is that a correction the control system can compute may be one the actuator cannot deliver without an unreasonable power bill. Section 1.4.5 of Chapter 1 sets out the same argument on a plate of five zones in a row.

This chapter works with a hundred-zone bake plate simulator on which the spatial frequency response, the boundary behavior, the wafer-drop transient and the per-zone run-to-run loop can each

be measured, and it prices four different disturbances in nanometers against the four different control elements that answer them. The simulator carries a 201-state thermal and chemical model with its instrument chain, three control architectures and two feedforward paths, and eleven scenarios that run from one power setting for a hundred zones to a graded lot on which the temperature map is perfect and the wafers are not. The thermal model, the spectral argument and the run-to-run arithmetic are standard. Section 5.2 traces how bake plates became a control problem. Section 5.3 gathers every equation and constant of the plate, the wafer and the resist, and Section 5.4 reads the spectrum of the plate and the three different gains a single loop can be said to have. Section 5.5 describes the zone loops, the three architectures, the wafer-drop pulse with its integral hold and the hundred run-to-run channels. Section 5.6 reports what each scenario measures, from the uniform actuator to the graded challenge. Section 5.7 names the thermocouples, heaters and controllers behind the loops. Section 5.8 states the limitations, Section 5.9 discusses plate design and the methods the model admits, and Section 5.10 summarizes.

5.2 Background

Resist bake was not always a control problem. A diazonaphthoquinone novolac resist prints a linewidth set mainly by exposure dose and develop time, and the post-exposure bake that follows exposure serves to relax standing waves rather than to drive the reaction that defines the feature. Bake plates of that era were single-zone aluminum slabs with one thermocouple and one cartridge heater, and the across-plate uniformity they achieved was a property of the metal and of the insulation around it rather than of any controller. Nothing in the tool was arranged to produce a deliberate temperature profile, because no profile was wanted.

Chemical amplification changed the arithmetic. Designing a resist so that one photogenerated acid catalyzes many deprotection events buys the sensitivity that short-wavelength exposure tools require [2], and it also moves the reaction that sets the linewidth out of the exposure step and into the bake [3]. Linewidth then acquires a bake-temperature sensitivity of several nanometers per degree, a coefficient later measured directly on extreme ultraviolet resists by holding a plate at a sequence of temperatures and reading the resulting critical dimension [4]. A single-zone plate that is flat to a degree is a plate that prints a range of several nanometers.

> ⧖ HISTORY **From Laboratory Curiosity to Production Resist (1980s)**
>
> Chemically amplified resists were proposed in the early 1980s, when exposure tools were moving into the deep ultraviolet and their lamps gave too little light there for the resists of the day. The first patent disclosure was rated an interesting laboratory curiosity and not worth filing. Yet the resist went quickly into the production of one-megabit memory chips, and it was so sensitive that the tool’s stage could not move fast enough: a filter was put in the beam to slow the exposure down. A serious problem followed. Airborne bases, at less than ten parts per billion, neutralized the acid while exposed wafers waited for their bake, and a skin formed on the developed lines. Filtering the air around the wafers through activated carbon became common practice, and a later resist that could be baked near 150 °C was barely affected by a long delay between exposure and bake [8].

The engineering answer arrived from two directions at once. Multizone temperature control had already been developed for rapid thermal processing, where lamp banks heat a wafer that has no thermal contact with anything, and where the modeling, identification and multivariable control of the resulting array were worked out in the early 1990s [5]. Applying the same idea to a bake plate is easier in the dynamics and harder in the specification, and the decentralized question of whether independent single-loop controllers suffice on such an array was posed in that literature [6]. Programmable thermal processing modules with many independently driven zones under a photoresist wafer followed [9], and zone-controlled post-exposure bake was demonstrated as a way of removing a global critical dimension signature on production mask blanks [10]. Later work closed the loop

Table 5.1 Variables of the Hundred-Zone Plate. Nominal Entries Describe the Design Point Loaded by Every Scenario.

Symbol	Role	Nominal	Range
T_i, $i = 1..100$	CV, measured	110 °C	90–130 °C SP
CD mean	CV, measured	62 nm	52–74 nm SP
P_i, $i = 1..100$	MV	30.80 W	0–120 W
ΔT_i^{SP}	MV, outer loop	0 °C	±4 °C
T_a ambient	DV, measured	22 °C	15–30 °C
T_{in} wafer in	DV, measured	22 °C	15–60 °C
c_d contact	DV, measured	1×	0.3–2×
b wafer bow	DV, unmeasured	0	0–0.8
s_x zone coupling	DV, measured	1×	0.2–3×
s_ℓ plate loss	DV, measured	1×	0.5–2×
g_h heater supply	DV, measured	1×	0.7–1.2×
g_f zone 34 heater	DV, unmeasured	1×	0–1×
β_0 TE bias	DV, unmeasured	0 °C	±2 °C
T_i	state (100)	110 °C	—
$T_{w,i}$	state (100)	22 °C	—
T_{rim}	state	98.42 °C	—

on a measurement of the wafer itself rather than of the plate, using in-situ sensing to control the steady-state spatial temperature profile [11], and multiplexed model predictive control was applied to the same multizone thermal plant when the computational cost of solving all zones at once became the binding constraint [12].

The correction that closes on the product came last, because it had to wait for the data infrastructure. Run-to-run control adjusts a recipe between wafers or lots using measurements taken after the step (Section 1.4.3), and it became standard practice in semiconductor manufacturing once advanced process control systems existed to carry a measurement from a metrology station back to the tool [13]. Extending a single run-to-run channel to an across-wafer map, so that a spatial signature rather than a mean is corrected, is a distinct step and has its own literature [14]. A hundred-zone plate is what an across-wafer run-to-run loop needs on the actuator side, and the reason zone counts rose is that lithographic scaling tightened the linewidth budget faster than the plates improved [15].

5.3 Process Model

Plate BP-301 is a 300 mm post-exposure bake plate cut into $N = 100$ heater zones, one under each of the hundred exposure fields of the wafer it bakes. Each field is 22 mm by 24 mm and the outermost field center sits at 130.2 mm from the wafer center, inside the 150 mm wafer radius. Zones are numbered on the field lattice, so a zone has four nearest neighbors unless it is on the boundary of the pattern, in which case the missing links attach instead to a passive plate rim that carries its own heat capacity and its own loss to ambient. Sixty-nine zones have four neighbors, eighteen have three and thirteen have two, and the forty-four missing links are rim links. Figure 5.1 draws the arrangement, the thermal paths and each symbol carried by the equations below.

The quantities the control system touches are sorted in Table 5.1. The plate carries 201 states, one temperature for each zone, one for the wafer element above it, and one for the rim. Working units are seconds, degrees Celsius, watts and nanometers, and the Arrhenius expressions convert to kelvin internally. Two clocks run above the physics, a zone loop that executes every 0.1 s and a run-to-run loop that moves once per wafer, which is every 110 s.

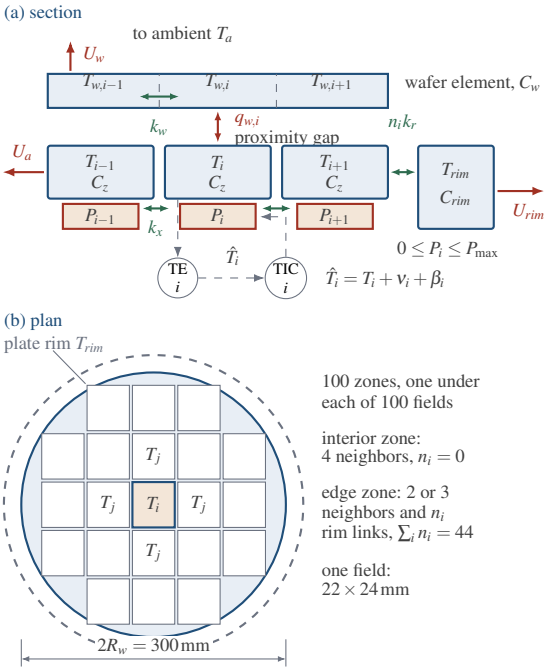


Figure 5.1 Bake plate BP-301. Section (a) shows three of the hundred cells, each a heater delivering P_i into a plate zone of heat capacity C_z at T_i, losing U_a to ambient, exchanging k_x with each neighbor and $n_i k_r$ with the rim, and passing $q_{w,i}$ across the proximity gap into the wafer element of heat capacity C_w at $T_{w,i}$, which conducts k_w sideways and loses U_w from its top face. Thermocouple TE-i reports $\hat{T}_i$ and TIC-i trims the feedforward power. Plan (b) shows the hundred fields on the wafer and the rim that the boundary zones are attached to.

5.3.1 Plate Zones, Rim and the Graph Laplacian

Each zone is a single lumped capacity driven by its heater and coupled to its neighbors, its rim links and the wafer above it, as shown in Equation (5.1).

$$\begin{aligned} C_z \frac{\mathrm{d}T_i}{\mathrm{d}t} &= P_i - U_a s_\ell (T_i - T_a) - q_{w,i} \\ &\quad - \sum_{j \in \mathcal{N}(i)} k_x s_x (T_i - T_j) \\ &\quad - n_i k_r s_x (T_i - T_{rim}) \end{aligned} \tag{5.1}$$

where C_z is the heat capacity of one zone in J/K, P_i is the delivered heater power in W, U_a is the loss to ambient, k_x is the conductance to one neighboring zone, $\mathcal{N}(i)$ is the set of lattice neighbors that exist, n_i is the number that do not, k_r is the conductance of one rim link, $q_{w,i}$ is the heat drawn by the wafer, and s_ℓ and s_x are operator multipliers on loss and on zone coupling. Both sides carry watts. The rim is one more lumped capacity collecting every rim link, as shown in Equation (5.2).

$$C_{rim} \frac{\mathrm{d}T_{rim}}{\mathrm{d}t} = \sum_{i=1}^{N} n_i k_r s_x (T_i - T_{rim}) - U_{rim} (T_{rim} - T_a) \tag{5.2}$$

With every zone at 110 °C and ambient at 22 °C, setting the left side of (5.2) to zero gives $T_{rim} = 98.42$ °C, and every scenario starts the rim state there so that the plate opens at a genuine equilibrium rather than climbing into one.

Collecting (5.1) at steady state over all hundred zones writes the plate as a matrix, as shown in Equation (5.3).

$$\mathbf{A}\,\theta = \mathbf{p}, \qquad \mathbf{A} = U_a\mathbf{I} + k_x\mathbf{L} + k_r\mathbf{D}_n \tag{5.3}$$

where θ holds the zone temperatures above ambient, $\mathbf{p}$ holds the powers, $\mathbf{L}$ is the graph Laplacian of the field lattice and $\mathbf{D}_n$ is the diagonal matrix of rim-link counts. Equation (5.3) is the entire control-relevant description of the plate. It is the two-dimensional form of the conductance matrix of Equation (1.8), with a rim in place of insulated ends, and Section 5.4 reads its spectrum.

5.3.2 The Wafer and Its Contact

A wafer is present only during the bake. While it is on the plate each wafer element exchanges with the zone below it, conducts to the wafer elements beside it and loses from its top face, as shown in Equation (5.4).

$$C_w \frac{\mathrm{d}T_{w,i}}{\mathrm{d}t} = q_{w,i} - U_w(T_{w,i} - T_a) - \sum_{j \in \mathcal{N}(i)} k_w(T_{w,i} - T_{w,j}) \tag{5.4}$$

$$\begin{aligned} q_{w,i} &= h_w\, c_d\, m_i\, (T_i - T_{w,i}) \\ m_i &= \mathrm{clip}\big(1 - b\, r_i^2 / R_w^2,\ 0.05,\ 1.5\big) \end{aligned} \tag{5.5}$$

where C_w is the heat capacity of one wafer element, h_w is the contact conductance of one cell, c_d is an operator contact multiplier, m_i is the contact map produced by a wafer bow b at field radius r_i, k_w is the lateral conductance across the silicon, and U_w is the loss from the free surface. Equation (5.5) is the only path by which a wafer defect enters, and it is deliberately radial, because a bowed wafer touches well near the center and badly at the edge. Off the plate the wafer element is not physically present, and the state is held inert at the incoming wafer temperature by $\frac{\mathrm{d}T_{w,i}}{\mathrm{d}t} = (T_{in} - T_{w,i})/\tau_h$ with $\tau_h = 2\,\mathrm{s}$, so that the state stays defined between bakes and is re-initialized to T_{in} the instant the next wafer lands.

Two numbers from Table 5.2 set the size of every result that follows. The hundred wafer elements total 89 J/K against 800 J/K for the hundred plate zones, and the wafer time constant $C_w/h_w = 2.54\,\mathrm{s}$ is half the plate zone time constant $C_z/(U_a + 4k_x) = 5.16\,\mathrm{s}$. In scenario 5 of Section 5.6 the wafer therefore covers 95 % of the 88 °C step from ambient to the plate setpoint 8.85 s after it lands and then rides 1.48 °C below the plate for the rest of the bake, which is the algebraic balance $U_w(T_i - T_a)/(h_w + U_w) = 1.483\,^{\circ}\mathrm{C}$ between the top-face loss and the contact conductance.

5.3.3 Thermal Dose and Linewidth

The resist integrates a reaction rate rather than a temperature, and because the rate is exponential in inverse absolute temperature, a cold interval and a hot interval of equal size do not cancel. The accumulated dose and the linewidth it produces are shown in Equations (5.6) and (5.7).

$$\Theta_i = \frac{1}{t_b}\int_0^{t_b} \exp\left[-\frac{E_a}{R}\left(\frac{1}{T_{w,i}} - \frac{1}{T_0}\right)\right] dt \tag{5.6}$$

$$CD_i = CD_0 - \kappa \ln\left(\frac{\Theta_i}{\Theta_0}\right) \tag{5.7}$$

where t_b is the bake time, E_a/R is the activation temperature of the deprotection in kelvin, $T_{w,i}$ and T_0 are absolute temperatures, κ is the linewidth change per unit of log dose, CD_0 is the linewidth of the calibrated bake and Θ_0 is the dose that calibrated bake delivers. The integral in (5.6) is accumulated at the 0.05 s physics step rather than at the controller scan, because it is the product.

The reference Θ_0 deserves its own sentence, because it is a calibration constant and not a derived one. A bake recipe is calibrated by printing wafers and measuring them, so the warm-up transient at the start of the bake, the landing pulse of Section 5.5 and the 1.48 °C by which the wafer sits below the plate are all inside the reference. Here $\Theta_0 = 0.917514$ is the retained historical reference for a

bake calibrated with the landing pulse running. It corresponds to holding the wafer isothermally at 108.95 °C for the whole 90 s. Writing (5.7) against Θ_0 rather than against unity is what maps a dose equal to the reference to 62.00 nm by construction, and it is also what makes every subsequent result a departure from a calibration rather than from an ideal. The reference is not recalibrated after correcting the six-second pulse clock. The shipped pulse prints 62.110 nm. A wafer held at its steady 108.52 °C for the whole bake, with no landing transient at all, would print 63.39 nm, so a bake without the pulse cannot reach the target without a temperature offset.

Differentiating (5.6) and (5.7) for a small uniform temperature offset held through the bake gives the sensitivity the whole plate specification is written from, as shown in Equation (5.8).

$$\frac{\mathrm{d}CD}{\mathrm{d}T} = -\frac{\kappa E_a / R}{T^2} = -\frac{39.1 \times 12000}{383.15^2} = -3.196\,\mathrm{nm/K} \tag{5.8}$$

A tenth of a degree is therefore 0.32 nm, which is a quarter of the 1.2 nm across-wafer range this application allows and a sixth of the 2.0 nm specification half-width on the mean.

5.3.4 Instruments

Each zone carries one thermocouple, and the controller sees $\hat{T}_i = T_i + \nu_i + \beta_i$, where ν_i is a Gaussian draw of standard deviation σ_T redrawn every controller scan and β_i is a fixed bias applied to zone 0 only. Every scenario except the graded challenge sets σ_T to zero. Critical dimension is measured on every field of every wafer at the end of the bake as $\widehat{CD}_i = CD_i + \sigma_{CD}\,\varepsilon_i$ with ε_i an independent standard normal draw. The instruments carry no dead time and no filter, and the run-to-run loop sees the map of the wafer that has just finished with no metrology queue in front of it.

5.3.5 Constants, Initial Condition and Numerics

Table 5.2 holds the constants. All but one of them are visible and adjustable in the model panel. The reference bake temperature T_0 in (5.6) is hidden, because a reader who can read it can read the answer to the dose calibration straight off the panel.

Every scenario opens with all hundred zones at the plate setpoint, the wafer elements at ambient, the rim at the equilibrium of (5.2), and each zone loop already holding the power the forward model of Section 5.5 computes. A startup test confirms that no state derivative at that condition exceeds 10^{-9} K/s. Fourth-order Runge-Kutta advances the 201 states with the step held at 0.05 s, the zone loops fire on a 0.1 s scan, and every logged channel reaches the historian five times per simulated second. The wafer cycle is counted in zone-loop scans, with its running-sum clock compared against its limits to within a microsecond, so the first wafer lands at $t = 20$ s, each bake lasts 90 s, the cycle repeats every 110 s and the shipped 6 s landing pulse runs for 60 scans. Each random draw comes from a mulberry32 stream seeded at 20260917, and the engine, its clocks and its seeding are those of Section 1.6.1, so a noisy scenario repeats draw for draw at any playback speed.

5.3.6 Assumptions

Detail was spent where a control decision depends on it and withheld everywhere else. Every simplification listed here has a condition under which it ceases to hold.

Two lumped layers per cell. A production plate has a heater layer, a plate body and a proximity air gap held by pins, and the gap contributes most of the dynamics as well as most of the sensitivity to wafer shape. Collapsing the gap into the single conductance h_w of (5.5) makes the landing transient first order when a real one is closer to second order, and it removes the difference between a wafer that is bowed and a wafer that is sitting on a particle.

Nearest-neighbor conduction with one isotropic coefficient. Production plates are cut with structured slots whose conductance is a deliberate design variable and is usually different along the two lattice directions [9]. The isotropic k_x in (5.1) makes the spectral gain of Section 5.4 separable, which is convenient and is not what a real plate does.

Table 5.2 Model Constants. All but the Reference Bake Temperature Are Shown and Editable in the Model Panel of the Application.

Symbol	Meaning	Value	Unit	Visibility
C_z	heat capacity, one zone	8	J/K	adjustable
U_a	zone loss to ambient	0.35	W/K	adjustable
k_x	conductance to a neighbor	0.30	W/K	adjustable
k_r	conductance of one rim link	0.15	W/K	adjustable
C_{rim}	heat capacity of the rim	240	J/K	adjustable
U_{rim}	rim loss to ambient	1.0	W/K	adjustable
$P_{\max}$	heater power per zone	120	W	adjustable
C_w	heat capacity, wafer element	0.89	J/K	adjustable
h_w	wafer contact conductance	0.350	W/K	adjustable
k_w	conductance across the wafer	0.10	W/K	adjustable
U_w	wafer top-face loss	0.006	W/K	adjustable
E_a/R	activation temperature	12000	K	adjustable
κ	CD per unit log dose	39.1	nm	adjustable
CD_0	CD of the calibrated bake	62	nm	adjustable
Θ_0	dose of the calibrated bake	0.917514	—	adjustable
—	CD specification half-width	2.0	nm	adjustable
—	across-wafer range allowed	1.2	nm	adjustable
t_b	bake time	90	s	adjustable
—	wafer exchange time	20	s	adjustable
σ_T	thermocouple noise, 1σ	0.02	°C	adjustable
σ_{CD}	CD-SEM repeatability, 1σ	0.40	nm	adjustable
T_0	reference bake temperature	110	°C	hidden

The wafer is in contact from the instant it lands. A real wafer is lowered onto proximity pins over roughly a second and the gap closes progressively, so the true landing transient is softer at its start than the step in (5.5) makes it. The assumption overstates how fast the first seconds of dose accumulate and therefore how sharp the feedforward pulse has to be.

One Arrhenius reaction stands for the resist. There is no acid diffusion length in (5.6), and acid diffusion is the physical effect that ultimately limits how spatially local any linewidth correction can be [3]. The model will therefore reproduce a zone-to-zone linewidth step that a real resist would blur.

The rim is one lumped state with no guard heater. Production plates carry a guard heater on the rim, which is itself a control loop and would remove most of the edge effect this application uses to make its point [9]. Raising U_{rim} or lowering k_r in Table 5.2 moves the plate toward that case.

Every field of every wafer is measured, with no queue. Production sampling plans measure a handful of sites per wafer and the measurement arrives one or more lots later, both of which make the run-to-run noise problem worse and the spatial filter of Section 5.5 more valuable rather than less [13]. Section 1.4.4 treats the metrology delay this assumption leaves out, and Chapter 6 runs a run-to-run loop through such a queue.

Heater power is continuous and instantaneous. Real zones are driven by solid-state relays with a cycle time of a second or more, which quantizes the correction and interacts with the loop gain.

5.4 The Spectrum of the Plate

5.4.1 Gain as a Function of Spatial Wavenumber

Away from the boundary the lattice in (5.1) is spatially invariant, so its steady-state response to a power pattern is diagonal in spatial frequency [7]. Take the interior of the plate with the wafer absent, write $\theta_i = T_i - T_a$, and substitute a plane wave $\theta_i = \hat{\theta}\, e^{\jmath \mathbf{q}\cdot\mathbf{x}_i}$ driven by $P_i = \hat{P}\, e^{\jmath \mathbf{q}\cdot\mathbf{x}_i}$ on an integer lattice

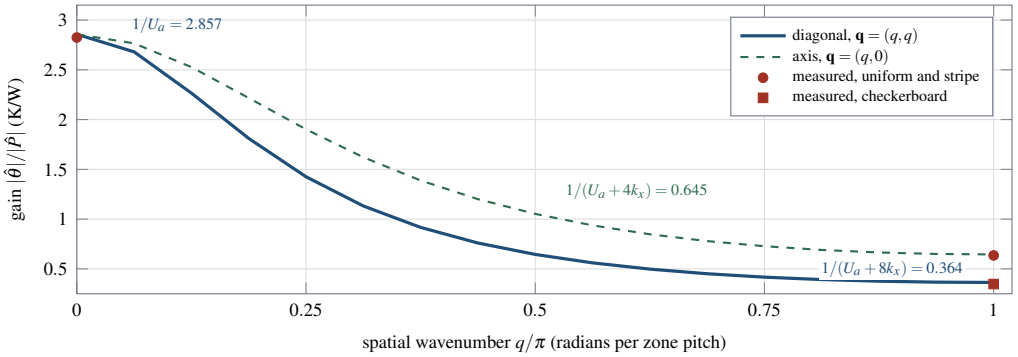


Figure 5.2 Steady-state gain of the plate against the spatial wavenumber of the power pattern, from Equation (5.9) with $U_a = 0.35\,\mathrm{W/K}$ and $k_x = 0.30\,\mathrm{W/K}$. The thick solid curve runs along the diagonal and the dashed curve along a lattice axis. The markers are steady states measured on the hundred-zone plate for a uniform 1 W addition, a striped $\pm 1\,\mathrm{W}$ pattern and a checkerboard $\pm 1\,\mathrm{W}$ pattern, averaged over the sixty-nine interior zones, filled circles for the uniform and striped patterns and a filled square for the checkerboard. The measured points sit slightly below the infinite-lattice curve because the finite plate loses heat to its rim.

$\mathbf{x}_i$. The neighbor sum collapses because $\sum_{j \in \mathcal{N}(i)} (\theta_i - \theta_j) = \theta_i\,(4 - 2\cos q_x - 2\cos q_y)$, and the result is Equation (5.9).

$$G(\mathbf{q}) = \frac{\hat{\theta}}{\hat{P}} = \frac{1}{U_a + 2k_x\,(2 - \cos q_x - \cos q_y)} \tag{5.9}$$

where q_x and q_y are spatial wavenumbers in radians per zone pitch and $\jmath$ is the imaginary unit. It is the two-dimensional counterpart of the mode gains of Equation (1.9). The number of zones does not appear in (5.9). What appears is the ratio of the diagonal loss to the coupling, and the three values a reader needs are the uniform pattern at $\mathbf{q} = 0$, the striped pattern at $\mathbf{q} = (\pi, 0)$ and the checkerboard at $\mathbf{q} = (\pi, \pi)$, as shown in Equation (5.10).

$$\begin{aligned} G(0,0) &= \frac{1}{U_a} = 2.857\,\mathrm{K/W} \\ G(\pi,0) &= \frac{1}{U_a + 4k_x} = 0.6452\,\mathrm{K/W} \\ G(\pi,\pi) &= \frac{1}{U_a + 8k_x} = 0.3636\,\mathrm{K/W} \end{aligned} \tag{5.10}$$

The ratio $G(0,0)/G(\pi,\pi) = 7.857$ is the spatial frequency response of the plate reduced to one number. It is a ratio of steady-state gains and not a bandwidth in time, so nothing about it changes when the loops are made faster. Figure 5.2 plots (5.9) along the two principal directions with the corresponding steady states measured on the simulator.

Adding 1 W to every zone of the simulated plate raises the sixty-nine interior zones by 2.8245 K on average (the whole plate by 2.7900 K) and the center zone by 2.8535 K, against 2.857 K from (5.10). A striped $\pm 1\,\mathrm{W}$ pattern produces 0.6353 K of amplitude against a predicted 0.6452 K, and a checkerboard produces 0.3476 K against 0.3636 K. The residual in each case is the boundary, which (5.9) does not model and the simulator does.

🕮 **Pause.** Try Exercise 5.1 before reading on.

5.4.2 Three Different Single-Loop Gains

The plate offers a reader three different answers to the question of what one loop's process gain is, and they differ by a factor of four. Holding the other ninety-nine zone temperatures fixed and adding

1 W to zone i moves only T_i, and the gain is the reciprocal of the diagonal of $\mathbf{A}$ in (5.3), which for an interior zone is $1/(U_a + 4k_x) = 0.6452\,\mathrm{K/W}$. That is the gain a loop sees when its neighbors are in perfect regulation, and it is the correct gain for decentralized tuning. Adding 1 W to zone i with every other heater left alone gives the lattice Green's function, measured on the simulator as 0.8005 K/W on the zone itself and 0.2006 K/W on each of its four neighbors. Adding 1 W to all hundred gives 2.857 K/W.

The first two numbers are exactly what the relative gain array is built from [1]. Since $\mathbf{A} = \mathbf{G}^{-1}$, the diagonal element is $\lambda_{ii} = G_{ii}A_{ii} = 0.8005 \times 1.55 = 1.241$ for an interior zone, which is close enough to unity that the array recommends the natural one-to-one pairing without reservation. That recommendation is correct and it is also not the whole story, because λ_{ii} compares the single-loop gain against the gain with every other loop closed, and it says nothing about the direction in which every loop moves together. The ratio of the common-mode gain to the single-loop gain is $G(0,0)/G(\pi,0) = 4.43$, and Section 5.6.2 measures what that factor does to a controller tuned on one loop. The value 1.241 belongs to an interior zone. A boundary zone has a smaller A_{ii} because it has fewer neighbors and a rim link instead, so its relative gain differs, and the array is not uniform across the plate even though every diagonal element stays near unity.

5.5 Controller Design

5.5.1 One Hundred Zone Loops

Each zone runs the velocity-form PID algorithm of Section 1.6.2. The derivative gain is fixed at zero in this app, so each zone loop is a PI controller, because the measurement is the plate itself and the useful anticipation arrives from the feedforward instead. In the zone's own symbols the increments are Equation (5.11), and the trim is assembled and limited as shown in Equation (5.12).

$$
\begin{aligned}
e_{i,k} &= a\,(T_{i,k}^{\mathrm{SP}} - \hat{T}_{i,k}), \qquad a = +1 \\
\Delta u_{i,k}^{P} &= K_c e_{i,k} - (K_c e_i)_{k-1}\,, \qquad \Delta u_{i,k}^{I} = \frac{K_c}{\tau_I}\, e_{i,k}\,\Delta t_c
\end{aligned}
\tag{5.11}
$$

$$
\begin{aligned}
s_{i,k} &= s_{i,k-1} + \Delta u_{i,k}^{P} \\
c_{i,k} &= s_{i,k} + I_{i,k-1} \\
u_{i,k} &= \mathrm{clip}\Big(c_{i,k},\ -P_{i,k}^{ff},\ P_{\max} - P_{i,k}^{ff}\Big) \\
I_{i,k} &= I_{i,k-1} + \Delta u_{i,k}^{I} + \frac{\Delta t_c}{T_t}\,\big(u_{i,k} - c_{i,k}\big)
\end{aligned}
\tag{5.12}
$$

where T_i^{SP} is the zone setpoint, $\hat{T}_i$ is the thermocouple reading, s_i is the running sum S_k of Equation (1.13), c_i is the trim the algorithm demands (v_k of Equation (1.14)), u_i is the trim the loop contributes in watts, I_i is the integral state of Equation (1.15), P_i^{ff} is the feedforward power described below, $\Delta t_c = 0.1\,\mathrm{s}$ is the scan period, and $T_t = \tau_I$ is the tracking time of the back-calculation, the routine's default because the app passes none. A zone reading above setpoint receives less power, so $a = +1$ in (5.11) is a reverse-acting loop in the convention of Section 1.6.2.

Two details of (5.12) are specific to this plant. The trim limits are the real heater limits less whatever the feedforward is already asking for, so the unwinding happens against the physical limit rather than a notional one. The back-calculation removes $\Delta t_c/T_t$ of the excess per scan (1 % at the default $\tau_I = 10\,\mathrm{s}$), and the running sum $s_{i,k}$ is never truncated. In Manual mode, and while the landing pulse of Section 5.5.3 runs, the routine is handed the trim actually applied and sets $I_{i,k} = u_{i,k} - s_{i,k}$, so the integral tracks rather than integrates.

Manual mode applies one power to all hundred zones, which is the arrangement scenario 1 exists to discredit. The model term is not applied in Manual, but each loop tracks the operator power less the feedforward its architecture will add in Auto, so a transfer to either closed architecture hands

over exactly the operator power. After 30 s in Manual in scenario 2 the center zone moves from 31.56 W to 31.55 W at the first Auto scan, the edge zone to 31.57 W, and the plate mean peaks at 110.16 °C. Open loop carries no trim and applies its own uniform power. The reverse transfer applies the mean delivered power to every zone, which is bumpless in the mean only. The tuning constants are bounded to $0 \le K_c \le 60\,\mathrm{W/^\circ C}$ and $0.5\,\mathrm{s} \le \tau_I \le 200\,\mathrm{s}$, and the output carries no rate limit. The scorecard scores the two displayed zone loops against a 0.15 °C band and the linewidth loop against a 0.4 nm band.

5.5.2 Three Architectures and a Forward Model

The feedforward term in (5.12) is what distinguishes the three architectures the application offers. Open loop supplies the power an interior zone needs and nothing else, $P_i^{ff} = U_a(T_i^{\mathrm{SP}} - T_a)$, with the loops tracking rather than acting. One hundred independent PI loops set $P_i^{ff} = 0$ and leave the whole power to the trim. The third architecture computes the power the entire setpoint profile requires from the conductances of (5.1), as shown in Equation (5.13).

$$\begin{aligned} P_i^{ff} = {} & U_a s_\ell (T_i^{\mathrm{SP}} - T_a) \\ & + \sum_{j \in \mathcal{N}(i)} k_x s_x (T_i^{\mathrm{SP}} - T_j^{\mathrm{SP}}) \\ & + n_i k_r s_x (T_i^{\mathrm{SP}} - T_{rim}) \end{aligned} \tag{5.13}$$

with an additional $U_w(T_i^{\mathrm{SP}} - T_a)$ while a wafer is present. Equation (5.13) is the row of $\mathbf{A}$ belonging to zone i applied to the setpoint vector, which is to say it is $\mathbf{A}\boldsymbol{\theta}^{\mathrm{SP}}$ evaluated one row at a time, the feedforward of Section 1.4.5. It is a static decoupler, like the one Volume 6, Chapter 3 places on a two-by-two plant, and the point worth making about it is the direction in which it runs. Asking what temperature a power pattern produces requires $\mathbf{A}^{-1}$, a dense inverse of a hundred-by-hundred matrix that has to be recomputed whenever a conductance changes. Asking what power a temperature profile needs requires $\mathbf{A}$ itself, which is sparse with at most five nonzero entries per row, and the answer is one matrix-vector product with no inverse in it at all. A controller only ever asks the second question. The rim temperature in (5.13) is read from the plant rather than predicted, which is legitimate because the rim is a measured state and not a manipulated one.

5.5.3 The Wafer-Drop Pulse and the Integral Hold

A cold wafer arriving on a hot plate is a disturbance whose time of arrival and size are both known before it happens, which makes it an accounting problem rather than a control problem (Section 1.4.7 treats feedforward on a known event). The app answers it with a rectangular power pulse added to every zone, triggered by the wafer landing and not by the error it causes, as shown in Equation (5.14).

$$P_i^{pulse}(t) = \begin{cases} P_p, & 0 \le t - t_{land} < t_p \\ 0, & \text{otherwise} \end{cases} \tag{5.14}$$

with defaults $P_p = 26\,\mathrm{W}$ and $t_p = 6\,\mathrm{s}$. The delivered power is then assembled and limited as shown in Equation (5.15).

$$\begin{aligned} P_i^{cmd} &= \mathrm{clip}\left(P_i^{ff} + u_i + P_i^{pulse},\ 0,\ P_{\max}\right) \\ P_i &= \mathrm{clip}\left(P_i^{cmd}\, g_{f,i}\, g_h,\ 0,\ P_{\max}\right) \end{aligned} \tag{5.15}$$

where P_i^{cmd} is the power the control system asks the heater for, $g_{f,i}$ is the zone heater health, unity everywhere except on the faulted zone, and g_h is a supply multiplier. The two clips in (5.15) are separate on purpose. A control system cannot command a heater above its rating whatever the arithmetic in front of it produces, and a heater that has lost most of its element delivers less than it

was asked for. The lower limit of zero is the one-sided actuator. A resistive heater cannot remove heat, so a zone that needs to come down has no actuator at all and the whole profile has to be lifted to make room for it. The saturation flag the scorecard reports is raised when P_i^{cmd} sits at either limit or the trim in (5.12) is clipped, which is a statement about the command rather than about the power that reached the plate. On a zone whose heater has failed the two differ by the heater health, and it is the command that says the loop has run out of authority.

While the pulse of (5.14) runs, the zone loops stop integrating and track their last trim instead, using the same $I_{i,k}$ assignment that manual mode uses. That hold is not an implementation convenience. Without it each loop sees the pulse as an error of its own making, drives its integrator down during the six seconds the pulse is applied, and then holds the plate below setpoint for the remainder of the bake while the integrator recovers, which undoes more than the pulse gained (Section 5.6.4 measures it). A feedforward term the feedback fights is worth less than no feedforward at all, and holding the integrators through a transient whose size is already known is the standard remedy.

🕮 **Pause.** Try Exercise 5.3 before reading on.

5.5.4 One Hundred Run-to-Run Channels

Above the zone loops sits an integral controller with one channel per zone, updated once per wafer from the measured linewidth map, as shown in Equation (5.16).

$$\begin{aligned} \delta_i &= \lambda_z \left(\widehat{CD}_i - CD^{\mathrm{SP}} \right) / s_{CD} \\ \Delta T_i^{\mathrm{SP}} &\leftarrow \mathrm{clip}\left(\Delta T_i^{\mathrm{SP}} + \delta_i,\ -\Delta T_{\max},\ \Delta T_{\max} \right) \end{aligned} \tag{5.16}$$

where $\lambda_z = 0.50$ is the update weight, $s_{CD} = 3.196\,\mathrm{nm/K}$ is the magnitude of the sensitivity in (5.8) and converts a linewidth error into a temperature correction, and $\Delta T_{\max} = 4\,^{\circ}\mathrm{C}$ bounds the accumulated correction. While the loop is on, the zone setpoints in (5.11) are $T_i^{\mathrm{SP}} = T^{\mathrm{SP}} + \Delta T_i^{\mathrm{SP}}$. Equation (5.16) is one hundred independent discrete integrators, each the incremental form of Equation (1.3) with the process gain estimated as $-s_{CD}$, and every one of them is driven by its own independent measurement noise. A field printed wide raises its zone setpoint, so each channel is direct acting in the convention of Section 1.6.2.

An optional spatial filter averages each update with those of its lattice neighbors before it is accumulated, replacing δ_i in (5.16) by $\bar{\delta}_i$ as shown in Equation (5.17).

$$\bar{\delta}_i = (1 - w)\,\delta_i + \frac{w}{1 + |\mathcal{N}(i)|} \left(\delta_i + \sum_{j \in \mathcal{N}(i)} \delta_j \right) \tag{5.17}$$

with $w = 0.60$ by default. The accumulated map is never filtered, so the filter shapes each increment and not the correction already in place. The app reports the spatial roughness of the correction map as a scalar, defined in Equation (5.18) as the root-mean-square of the discrete Laplacian of the offset map divided by the root-mean-square of the map itself.

$$\begin{aligned} \mu_i &= \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \Delta T_j^{\mathrm{SP}} \\ c^2 &= \sum_i \left(\Delta T_i^{\mathrm{SP}} - \mu_i \right)^2 \Big/ \sum_i \left(\Delta T_i^{\mathrm{SP}} \right)^2 \end{aligned} \tag{5.18}$$

A smooth radial correction scores near zero on (5.18) and a correction that is independent from zone to zone scores near one. Equations (5.9) and (5.18) together say what the filter is for. A checkerboard offset of 1 °C amplitude costs $1/G(\pi,\pi) = 2.750\,\mathrm{W}$ per zone while a uniform offset of the same amplitude costs $1/G(0,0) = 0.350\,\mathrm{W}$, so every unit of roughness in the correction is paid for at 7.86 times the watts per degree of the part of the correction that was in the product to begin with.

Table 5.3 Scenarios and Completion Requirements

#	Scenario and kind	What it shows	Required to complete
1	One power for a hundred zones, Manual *experiment*	a uniform actuator on a plant with a boundary	In Manual, apply the uniform setpoint power (30.80 W) until the plate range exceeds 5.2 °C, then raise every zone to at least 31.5 W for 60 s and compare the shape of the map with its level.
2	Holding the profile *experiment*	common-mode against single-loop gain	After at least 2 s of running, step the plate setpoint to 110.5 °C with $K_c < 10\,\mathrm{W/°C}$ until the center and edge zones are within 0.05 °C of it, then reload the case, raise K_c to 20 W/°C, run 2 s at 110 °C and step again until both are within 0.05 °C.[b]
3	Open loop and the edge *experiment*	the forward model against feedback	Settle in open loop to a plate range above 5.2 °C and switch to a hundred PI loops until it is below 0.05 °C, then reload the case, settle again and reach the same range under Model plus PI.
4	A hundred independent loops, zone 34 heater fails at $t = 120\,\mathrm{s}$ *demonstration*	one fault read off two maps	Run a hundred PI loops at $K_c \geq 10\,\mathrm{W/°C}$ through the failure until a CD range above 10 nm, then Reset with $K_c \leq 5\,\mathrm{W/°C}$ and run through the failure until a wafer baked after it again shows more than 10 nm.
5	A wafer lands *demonstration*	a disturbance no loop is fast enough for	With the landing pulse off, see the plate mean fall more than 2 °C below setpoint during a bake, then finish that wafer.
6	The feedforward pulse *experiment*	timed feedforward and the integral hold	Finish a wafer with a pulse of 24 to 28 W, then Reset and bake a wafer at 16 W or less that prints above 64 nm and one at 38 W or more that prints below 60 nm.
7	A bowed wafer *demonstration*	a signature no plate instrument can see	See the plate range exceed 0.8 °C at a landing, then with run-to-run off and bow at least 0.3 bake three wafers until the CD range exceeds 1.2 nm while the plate range is below 0.3 °C.
8	Run-to-run, one channel per zone *experiment*	a hundred integrators and their noise	With run-to-run on, reach a CD range below 2.5 nm once three wafers are baked and, with the correction per zone, below 1.2 nm once six are.
9	The frequencies the plate cannot produce *experiment*	spatial bandwidth as a cost	With the spatial filter on, reach a CD range below 1.2 nm once seven wafers are baked, then set $w \geq 0.95$, fail the zone 34 heater and hold a CD range above 10 nm for 220 s.
10	A thermocouple reading high, $+0.5\,°\mathrm{C}$ at $t = 200\,\mathrm{s}$ *demonstration*	a loop that satisfies a bad instrument	After the bias, bake to the third wafer with a CD range above 1.0 nm, then switch run-to-run on and bake to the fifth with a zone 0 offset above 0.15 °C.
11	Challenge, a bowed lot *challenge*	four disturbances, four answers, in order	Over 700 s (six wafers), keep the mean CD within 0.35 nm of 62 nm from $t = 462\,\mathrm{s}$ (wafers 4 to 6), finish with an across-wafer range below 1.2 nm and a mean zone offset within $\pm 1\,°\mathrm{C}$, with the model parameters and disturbances as given.[a]

[a] As issued the worst mean-CD error over wafers 4 to 6 is 5.69 nm, sixteen times the tolerance, with a 3.12 nm range and all six wafers out of specification. The suggested set (Model plus PI, the pulse, $K_c = 6\,\mathrm{W/°C}$, $\tau_I = 12\,\mathrm{s}$, filtered run-to-run at $\lambda_z = 0.6$) gives 0.09 nm, 0.83 nm and 0.40 °C, passing with margins of 0.26 nm, 0.37 nm and 0.60 °C, identically after a Reset and on eight other noise seeds. [b] A setpoint chosen before the run starts is where the plate starts, so both steps are made on a running plate. Both zones settle within 0.05 °C 9.9 s after the step at $K_c = 4\,\mathrm{W/°C}$ and 0.75 s after it at 20 W/°C.

5.6 Results

Table 5.3 sets out the eleven scenarios in the order the app lists them, with what each one shows and what completes it (Section 1.6.4 describes scenarios and completion rules). Each one pins the mode, the architecture, the tuning, the disturbances and any timed event, and each puts its question to the reader ahead of the run rather than after it. Every result below was measured on the app's own engine with the case selected and any stated change made before the run starts, unless the text says otherwise. Table 5.4 holds what was measured, and Figure 5.3 redraws its wafer rows. Figure 5.4 shows the app during scenario 9. The exported file keeps the true temperature of the center zone beside its thermocouple reading, the mean and the extremes of the linewidth map, the spread of the zone offsets and the spatial roughness of the correction.

Table 5.4 Measured Outcomes Read From the App's Own Scorecard. Plate Range Is the Spread Across the Hundred Zones at the End of a Bake and CD Figures Are Those of the Last Completed Wafer. Rows 8 and 9 Report the Mean Over Wafers 10 to 20 of the Shipped Noise Seed and Challenge Rows the Sixth Wafer.

#	Configuration	Plate range (°C)	Mean *CD* error (nm)	*CD* range (nm)	Roughness *c* (—)	Note
1	MANUAL, 30.80 W every zone	5.437	—	—	—	17.38 nm of linewidth in the plate map
3	Open loop, same power	5.437	—	—	—	corner zones 104.45 °C, interior 109.89 °C
3	Forward model, all feedback removed*	1.1×10^{-12}	—	—	—	$K_c = 0$, no trim, from open-loop profile
2	Model plus PI, $K_c = 4$, $\tau_I = 10$ s	0.000	—	—	—	settles from the open-loop profile in 52.8 s
4	100 PI loops, heater fault, wafers baking	12.089	+4.609	23.635	—	faulted field 23.63 nm off the median
5	Model plus PI, wafer landing, no pulse	0.000	+4.015	0.010	—	plate mean dips 2.777 °C
6	Pulse 26 W for 6 s	0.001	+0.110	0.266	—	dip 0.096, overshoot 9.290 °C
6	Pulse 26 W for 2 s	0.000	+3.869	0.004	—	dip 0.287, overshoot 1.238 °C
6	Pulse 14 W for 6 s	0.001	+3.592	0.020	—	dip 1.539 °C
6	Pulse 40 W for 6 s	0.003	−4.928	0.798	—	dip 0.000 °C
7	Bowed wafers $b = 0.5$, no run-to-run	0.002	+1.996	3.765	—	radial, center 62.19 nm to edge 65.95 nm
8	Run-to-run per zone, $\lambda_z = 0.5$	1.702	+0.011	0.745	0.254	offset map RMS 0.785 °C, spread 1.71 °C
9	Run-to-run spatially filtered, $w = 0.6$	1.599	+0.014	0.595	0.189	offset map RMS 0.778 °C, spread 1.61 °C
10	Thermocouple on zone 0 reads +0.5 °C	0.500	+0.126	1.298	—	zone 0 held at 109.501 °C
11	Challenge, tuning as issued	0.008	+5.693	3.120	—	6 of 6 wafers out of specification
11	Challenge, suggested configuration	1.223	+0.009	0.832	0.234	1 of 6 wafers out, mean offset 0.40 °C

* A model run outside the app's controls. In the app, switching architecture carries the open-loop power into the trims, so with $K_c = 0$ the plate keeps its 5.437 °C profile, and scenario 2 at $K_c = 0$ holds its flat start exactly.

5.6.1 A Uniform Actuator on a Plant With a Boundary

Scenario 1 gives every zone the power an interior zone needs. That power is $U_a(T^{\mathrm{SP}} - T_a) = 0.35 \times 88 = 30.80$ W, and the interior zones farthest from the edge reach 109.89 °C because their neighbors are at nearly their own temperature and little heat flows sideways. The sixty-nine four-neighbor zones average 108.99 °C, the eighteen three-neighbor zones 106.18 °C and the thirteen two-neighbor corner zones 104.72 °C, the coldest at 104.45 °C, against a rim that has fallen to 94.36 °C. The plate range is 5.437 °C, which at the 3.196 nm/K of (5.8) is 17.38 nm of linewidth against a 1.2 nm across-wafer allowance. None of that is a control failure. It is what a uniform actuator does to a plant with a boundary, and it is why the zones exist. Raising every zone to 31.8 W lifts the plate mean from 107.93 °C to 110.72 °C and scales every temperature rise above ambient by the same factor, so the level of the map moves and its shape does not.

The power the edge zones need is not a mystery either. Evaluating (5.13) on a flat 110 °C profile returns 30.800 W for a four-neighbor zone, 32.537 W for a three-neighbor zone and 34.274 W for a two-neighbor corner, a total of 3156.4 W against the 3080 W the uniform recipe delivers. In a model run outside the application's controls, loading those powers with no feedback at all (loop gain zero, trims zero) brings the plate from the 5.437 °C profile to a range of 1.1×10^{-12} °C, which is the

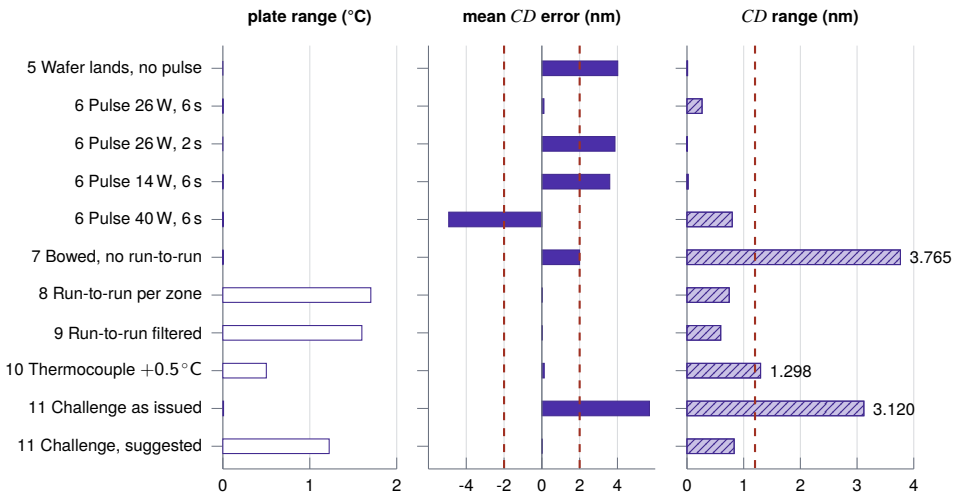


Figure 5.3 The wafer rows of Table 5.4 side by side: plate range at the end of a bake (open bars), mean linewidth error (solid bars, the 2.0 nm specification half-width dashed) and across-wafer linewidth range (hatched bars, the 1.2 nm allowance dashed). The rows of scenarios 1 to 3 carry no linewidth, and the heater fault of scenario 4 is left out because its 23.635 nm range would set the scale. Notice that the bowed wafers of scenarios 7 to 9 and 11 come into specification only on the three plates left least flat.

residual of the integrator and not of the model. In the application, scenario 2 at $K_c = 0$ holds its flat start exactly. The approach takes 56.9 s to a range below 0.05 °C, because the edge zones wait on the rim, whose time constant $C_{rim}/(U_{rim} + \sum_i n_i k_r) = 31.6\,\text{s}$ is six times that of a zone. Selecting either closed architecture from the settled open-loop plate carries each zone's command into its new trim and keeps the present proportional term in the loop's running sum, so the transfer is bumpless. The coldest corner zone holds 30.80 W through the scan that applies the change and moves 0.22 W on the next, and the plate mean never falls below its open-loop value. The plate then reaches a range below 0.05 °C in 44.7 s under a hundred PI loops and 52.8 s under model plus PI. The hundred PI loops are the faster of the two, which is the opposite of what a reader expects a decoupler to buy. At the instant of a bumpless switch the trims cancel the extra edge power the model asks for, and they release it only as fast as they integrate. Both beat the feedforward alone because the loops drive the corner zones above their final 34.32 W (35.24 W after 10 s) while the rim warms. Speed is not what the model buys. What it buys is a power profile computed once from the plate drawing rather than searched for by a hundred integrators, which leaves the authority those integrators would have spent available for the disturbances nobody wrote down.

5.6.2 A Gain Tuned on One Loop

Scenario 2 makes the spectral argument of Section 5.4 operational. In the application the plate runs Model plus PI and the common-mode step of 0.5 °C brings the center and edge zones within 0.05 °C of the new setpoint 9.9 s after a step made on the running plate at $K_c = 4\,\text{W}/^\circ\text{C}$, with a 10.6 % center overshoot, and 0.75 s after it at 20 W/°C, because the forward model carries the new power at once and the loops only trim it. A reader who tunes one zone by stepping that zone's setpoint alone measures the single-loop gain of 0.6452 K/W and a time constant of 5.16 s, and a SIMC tuning that asks for a closed loop no faster than the plant itself returns $K_c = 1/0.6452 = 1.55\,\text{W}/^\circ\text{C}$ and $\tau_I = 5.16\,\text{s}$ [16]. The same procedure run on the common mode measures 2.857 K/W and 22.86 s and returns $K_c = 0.35\,\text{W}/^\circ\text{C}$ and $\tau_I = 22.86\,\text{s}$. The two gains differ by exactly $G(0,0)/G(\pi,0) = 4.43$.

Table 5.5 The Same Two Tunings on Two Different Spatial Directions, Measured on One Hundred PI Loops With No Feedforward. A 0.5 °C Setpoint Move Is Applied to One Zone or to All Hundred, and the Center Zone Is Read. Settling Is Timed From the Scan at Which the Move Takes Effect.

	single-loop tuning $K_c = 1.55$, $\tau_I = 5.16$		common-mode tuning $K_c = 0.35$, $\tau_I = 22.86$	
Move	one zone	all zones	one zone	all zones
Overshoot (%)	0.0	18.8	0.0	0.0
Peak rise (°C)	0.500	0.594	0.421	0.500
Settling to ±0.02 (s)	15.35	23.55	> 200	73.35
Loop gain $K_c G$ (—)	1.00	4.43	0.226	1.00

Table 5.5 applies each tuning to both kinds of move on the same plate with a hundred PI loops and no feedforward. It is a model experiment rather than a run in the app, because the app offers no setpoint for a single zone and its slider grid does not hold these exact constants. The nearest slider settings, $K_c = 1.6\,\mathrm{W/°C}$ and $\tau_I = 5\,\mathrm{s}$, overshoot 19.1 % on the all-zone move and settle in 22.8 s. The single-loop tuning is textbook on the move it was designed for and overshoots 18.8 % when all hundred zones step together, because the loop gain product on that direction is $K_c\,G(0,0) = 4.43$ instead of the unity the rule intended. The common-mode tuning reverses the failure exactly, reaching only 84 % of a single-zone step within 200 s because its loop gain product on that direction is 0.226. Neither controller is mistuned in any sense a single step test can reveal.

5.6.3 A Local Fault With a Distributed Signature

Scenario 4 bakes wafers with a hundred independent PI loops and removes 88 % of one heater's authority at $t = 120\,\mathrm{s}$, between the first and second wafers of the lot. It is worth reading twice, because what the plate does and what the wafer does are different, and the scenario is arranged so that the same fault can be read off both maps at once. On the plate the fault is one zone wide. Sixty seconds into any bake from the third wafer on, the faulted zone sits 12.072 °C below its setpoint while its three neighbors are within 0.003 °C of theirs, and exactly one zone of the hundred is more than 0.003 °C off. The neighbors succeed completely. Each of them sees heat leaving sideways, raises its own power, and holds its own zone, so the temperature map shows one cold square and nothing around it. The instant has to be named because the neighbor deviation is still decaying at that point, from 0.043 °C at 30 s into the bake through 0.0022 °C at 60 s to 0.0001 °C at the end of it. What the neighbors are finishing off is the wafer landing and not the fault, because the same decay repeats to the fourth decimal on every wafer from the third onward. What is diagnostic on the plate is not the temperature map but the power map and the saturation flag, because the faulted loop is commanded the full 120 W rating and delivers 14.40 W through the damaged element. The flag is raised at the scan 2.1 s after the fault lands and stays raised for the rest of the run, and it is clear in every other scenario as shipped. On the wafer the same fault is five zones wide. Silicon conducts sideways with nothing regulating it, so the wafer above the faulted zone runs 6.97 °C below the median of the eighty-six far-field elements and the wafer above the three neighbors runs 1.03 °C, 1.14 °C and 1.31 °C below it. Those three differ from each other because the three zones carry different nominal powers, 30.80 W, 32.54 W and 34.27 W, so the edge effect of Section 5.6.1 reappears inside the fault. The linewidth map that results puts the faulted field 23.63 nm off the median, its three neighboring fields 3.43 nm to 4.39 nm off and the second ring 0.52 nm to 1.08 nm off, with 11 of the hundred fields more than 0.2 nm from the median and 8 more than 0.5 nm. The spreading is the k_w term of (5.4) and not conduction in the plate. The width of the signature is set by the wafer and not by the actuator that caused it, which is what makes reading the map backward to find the broken heater a skill rather than a lookup. The first wafer of the lot is baked before the fault and is worth keeping. It comes out at 66.210 nm with an across-wafer range of 0.001 nm, which is the wafer-drop error of Section 5.6.4 arriving on its own

because this scenario runs with the feedforward pulse switched off. Every later wafer carries the same 66.21 nm median with the fault signature on top of it. Two faults of completely different kinds are therefore visible in the same map, one uniform and one local, and only the local one moves the range.

5.6.4 A Disturbance With a Known Arrival Time

Scenario 5 drops a 22 °C wafer onto a 110 °C plate with every loop healthy and the forward model in service. The plate mean falls 2.777 °C and is back within 0.1 °C 33.7 s after the landing, and the linewidth comes out 4.015 nm wide of target on every field of the wafer, which is twice the 2.0 nm specification from one event on a tool whose loops are all working. The across-wafer range is 0.010 nm, which is the point. Nothing about this failure is a uniformity failure.

> **TRY IT IN THE APP** **5 · A Wafer Lands**
>
> Load the scenario and press Run. Watch the landing strip as the first wafer lands: the plate mean falls 2.777 °C and is back within 0.1 °C 33.7 s later. When the bake ends, read the wafer note against the 62 nm target. The linewidth comes out 4.015 nm wide on every field, although every loop on the plate is working.

Most of that cost is not the dip. The wafer heats by taking heat out of the plate, so the plate is coldest exactly while the wafer is climbing, but the wafer is then far below the plate and the Arrhenius rate it integrates is small. Integrating the dose of (5.6) through one bake shows where the deficit lives. The first 10 s of the bake contribute 4.31 % of the final dose against the 11.11 % they would contribute at a constant rate, and by 30 s the accumulated dose has recovered to 28.53 % against 33.33 %. A 26 W pulse held for only 2 s removes the dip almost entirely (0.287 °C, with a 1.238 °C overshoot) and the line still prints 65.869 nm. The deficit is the calibration of Section 5.3.3, whose reference dose belongs to a bake with the pulse running. A simple mixing estimate $C_w^{tot}(T^{\mathrm{SP}} - T_{in})/(C_w^{tot} + C_z^{tot}) = 8.81\,^{\circ}\mathrm{C}$ badly overstates the dip, because the heaters keep running and the plate zone time constant is twice the wafer time constant, but it gives the right order.

> **DID YOU KNOW?** **The First Seconds on the Plate**
>
> In a 2005 experiment a wafer set on a bake plate climbed from 24 °C to 110 °C in 3.1 s. A thin polyimide sheet under the wafer stretched the climb to 30.5 s, and the printed lines came out wider by about 0.3 nm for each second of ramp [4].

Feedback cannot answer it. The shortfall is over in the time a filtered thermocouple takes to decide that it happened, and it has a known time of arrival and a known size. Scenario 6 announces it instead. Table 5.4 reports a sweep of the pulse amplitude at fixed length. At 26 W for 6 s the dip falls from 2.777 °C to 0.096 °C and the mean linewidth returns to 62.110 nm, an error of +0.110 nm. The plate does not stay flat to deliver it. The pulse puts 15.6 kJ into the plate (26 W on a hundred zones for 6 s) where the wafers need about 7.8 kJ, so the plate mean overshoots the setpoint by 9.290 °C and the wafer peaks at 114.11 °C. At 14 W the pulse leaves 3.592 nm on the table and at 40 W it overshoots by 4.928 nm, so the optimum is findable and it is not wide. The same values result when the amplitude is changed between wafers. The length sweep is sharper still, because at 26 W a pulse of 4 s leaves 2.490 nm and one of 8 s overshoots by 3.130 nm. Both curves are energy balances in dose rather than control problems, which is the lesson.

The part of the pulse that is not the pulse is the integral hold described in Section 5.5.3. It is switched on whenever the pulse is running and cannot be switched off in the application, so its value was measured on a copy of the controller with the hold removed. Without the hold each zone loop reads the 26 W as an error of its own making, and by the end of the pulse the center zone trim has fallen to −17.4 W against −0.8 W with the hold. The plate mean then peaks at only 113.72 °C, sags to 108.98 °C while the integrators climb back, and the wafer prints 66.450 nm, wider than the 66.015 nm of no pulse at all.

5.6.5 The Plate Is Not the Product

Scenario 7 uses wafers with $b = 0.5$ in (5.5), which touch the plate well at the center and badly at the edge. The linewidth map that results is a clean radial signature running from 62.19 nm at the wafer center to 65.95 nm at the outermost field, a range of 3.765 nm and a three-sigma spread of 3.253 nm. The plate range at the end of the bake is 0.0016 °C, which is 0.005 nm of equivalent linewidth. Every thermocouple on the tool reads its setpoint and the tool is printing three nanometers of across-wafer range.

> **TRY IT IN THE APP** **7 · A Bowed Wafer**
>
> Press Run and bake three wafers with Run-to-run zone offsets left Off. Watch the CD map build a ring pattern while the plate map stays flat. The linewidth runs from 62.19 nm at the center to 65.95 nm at the outermost field, a range of 3.765 nm, while the Plate range badge reads about 0.0016 °C at the end of each bake.

Only a measurement of the wafer can see that, which is the argument for run-to-run correction and not merely a preference for it [11, 14]. Scenario 8 turns Equation (5.16) on. The radial signature is removed in four or five wafers, with the across-wafer range falling from 3.764 nm to 1.074 nm by the fourth wafer and to a mean of 0.745 nm over wafers 10 to 20. The mean linewidth returns to 62.011 nm.

Pause. Try Exercise 5.2 before reading on.

The plate is no longer flat, and it is not supposed to be. Over the same wafers the zone offsets span 1.71 °C and the plate range at the end of a bake is 1.702 °C against the 0.002 °C of scenario 7. A temperature map read on its own would rank that plate below every healthy closed-loop case, and it prints wafers with a fifth of the range that the flat plate of scenario 7 prints.

5.6.6 Noise Filtered in Space Rather Than in Time

What also comes out of scenario 8 is roughness. Every channel of (5.16) is driven by its own independent 0.40 nm measurement draw, no two neighboring zones ever see the same number, and the offset map accumulates a grain that is a random walk with nothing to stop it. The roughness of (5.18) climbs from 0.213 after the first wafer to a mean of 0.254 over wafers 10 to 20, and over the same wafers the discrete Laplacian of the offset map has a root-mean-square of 0.1996 °C against a map root-mean-square of 0.785 °C.

The spatial bandwidth limit does not appear here as an inability. The forward model of (5.13) inverts the plate exactly and will ask for a checkerboard without complaint, and the plate will produce one. It appears as a cost, at 7.86 times the watts per degree that a smooth correction of the same amplitude would need, spent on a pattern that was measurement noise before it was a correction and that is printed on every wafer thereafter.

Scenario 9 applies the neighbor average of (5.17) before the correction reaches the setpoints. Over wafers 10 to 20 the roughness falls from 0.254 to 0.189, the Laplacian root-mean-square of the offset map from 0.1996 °C to 0.1473 °C, and the across-wafer linewidth range from 0.745 nm to 0.595 nm, with the three-sigma spread falling from 0.446 nm to 0.358 nm. The comparison holds across noise seeds. On five further seeds the mean range runs from 0.685 to 0.848 nm per zone and from 0.545 to 0.661 nm filtered, and the filtered run is lower on every seed by 0.11 to 0.24 nm. The magnitude of the correction itself is nearly untouched, at 0.785 °C and 0.778 °C root-mean-square respectively, and so is the speed of the initial convergence (1.074 nm against 1.196 nm at the fourth wafer and 0.984 nm against 0.943 nm at the fifth), because a radial signature carries little energy at the wavenumbers (5.17) removes. The filter costs nothing on the disturbance and takes a quarter out of the roughness, which is the whole argument for it. The Q-filter of the iterative learning controller in Volume 8, Chapter 10 plays the same part for a correction stored along a scan path.

The limit is equally real, and scenario 9 completes on it. When the zone 34 heater fails after the seventh wafer, no setting removes its field, because the zone is commanded at its 120 W rating and its offset reaches the 4 °C bound within one to three wafers. Six wafers later the CD range is 17.46 nm at $w = 1$, 17.60 nm at $w = 0.6$ and 18.33 nm per zone, against 24.29 nm with the run-to-run loop off. What the filter changes is where the rest of the correction lands. At $w = 1$ the three neighboring fields print 1.01 nm to 1.56 nm narrow and six fields besides the faulted one sit more than 0.5 nm from the median, against three fields, none more than 0.69 nm off, per zone. Spatial filtering trades resolution against noise in exactly the way temporal filtering trades speed against noise, and for exactly the same reason.

5.6.7 An Instrument That Lies

Scenario 10 biases the thermocouple on zone 0 by $+0.5\,^{\circ}\mathrm{C}$ at $t = 200\,\mathrm{s}$. The loop reduces the power of that zone at the end of a bake from 34.71 W to 34.03 W until the plate really is half a degree cold at 109.501 °C, and then reports that it is exactly on setpoint, a measurement that is wrong in the sense of Section 1.4.2. The linewidth on that field moves from 62.281 nm to 63.330 nm and the across-wafer range from 0.266 nm to 1.298 nm. The 1.050 nm of linewidth is smaller than the 1.598 nm that (5.8) predicts for half a degree, because the silicon above a single cold zone is being warmed from the side through k_w, which is the same mechanism that spread the heater fault. Turning the run-to-run loop on after the third wafer removes the error, because it measures the product rather than the plate, and from the eighth wafer on the field stays within 0.26 nm of the median. What it leaves behind is a permanent zone offset, 0.51 °C to 0.67 °C on zone 0 from the seventh wafer to the twenty-first, that is a record of an instrument fault nobody has diagnosed.

5.6.8 The Graded Challenge

The challenge presents four disturbances at once. The wafers are bowed at $b = 0.45$ and arrive at 19 °C, the forward model is switched off, the pulse is off, the run-to-run loop is off and the loops are slow at $K_c = 2\,\mathrm{W/^{\circ}C}$ and $\tau_I = 40\,\mathrm{s}$, with thermocouple noise of 0.02 °C. The target has three parts over the 700 s lot of six wafers: a mean linewidth within 0.35 nm of 62 nm over the last third (wafers 4 to 6), an across-wafer range under 1.2 nm on the last wafer, and a mean zone offset within $\pm 1\,^{\circ}\mathrm{C}$ at the end. As issued the configuration prints 67.693 nm with a range of 3.120 nm, and all six wafers of the lot carry fields outside the specification. The suggested configuration turns on the model, the pulse, the run-to-run loop at $\lambda_z = 0.6$ and the spatial filter, and raises the gain to $K_c = 6\,\mathrm{W/^{\circ}C}$ with $\tau_I = 12\,\mathrm{s}$. It prints 62.009 nm with a range of 0.832 nm by the sixth wafer and a mean zone offset of 0.398 °C, the worst mean error over wafers 4 to 6 is 0.087 nm, and only the first wafer of the lot (63.20 nm) is out of specification. On eight other noise seeds the same settings pass with worst errors of 0.08 to 0.14 nm and ranges of 0.59 to 0.85 nm.

Read over the same 700 s window, the linewidth loop integrated absolute error falls from 3395.3 nm s to 222.3 nm s and the peak linewidth error from 5.998 nm to 1.198 nm. The zone-loop numbers move the other way, because the center zone integrated absolute error rises from 349.5 °C s to 862.6 °C s and its total variation from 252.3 W to 852.3 W. The configuration that prints good wafers is the configuration whose temperature scorecard looks worse, since the pulse and the per-zone offsets both move the plate on purpose. Ordering matters as well. The run-to-run loop alone ($\lambda_z = 0.6$, filtered) brings the sixth wafer to 62.040 nm with a 0.76 nm range, but only by writing the edge and the landing into the zone offsets, which end at a mean of 1.815 °C, and the offset part of the target rejects it on every seed tried. Switching the run-to-run loop on first and the pulse two wafers later counts the landing twice, so the next wafer prints 58.52 nm, and the integrators of (5.16) hand the excess back only one wafer at a time (60.52, 61.34 and 61.725 nm), too late for the last third of the lot. The pulse itself is sized for tuned loops. On the slow loops as issued it over-bakes the line to 55.27 nm, and with the model at $K_c = 6\,\mathrm{W/^{\circ}C}$, $\tau_I = 12\,\mathrm{s}$ it gives 63.2 nm, which leaves the last nanometer and the bow to the run-to-run loop.



Figure 5.4 The hundred-channel run-to-run loop of scenario 9 at $t = 21.67$ min, correcting bowed wafers with the correction spatially filtered, cropped to the plant view and the trends. The left map (1) is the plate temperature about its setpoint and the right map (2) is the linewidth the resist produced from it, drawn on the same hundred cells, with the power of every zone as a bar field underneath (3). A twelfth wafer is 70 s into its 90 s bake and the map on the right belongs to wafer 11, whose across-wafer linewidth range has fallen to 0.56 nm against a 1.20 nm allowance at a mean of 62.01 nm. The plate is deliberately 1.517 °C out of flat because that is the profile the linewidth map asked for, the mean zone offset has reached 0.652 °C and carried the rim with it to 100.0 °C, and the spatial roughness of the correction is holding at 0.188. The trends show the whole run on a 30 min axis, with one heater pulse per bake on the two zone traces (4) and a linewidth mean stepping down toward its target once per wafer. The lowest strip in frame (5) collects the across-wafer numbers on one axis, the linewidth range against its allowance, the plate range, the spread of the zone offsets and the spatial roughness. An eighth strip carrying the wafer-landing detail sits below the visible area.

✦ DID YOU KNOW? A Flat Plate and a Failed Lot

From the simulator: in Table 5.4 the challenge as issued ends with a plate range of 0.008 °C and six of six wafers out of specification. The suggested configuration ends with a plate range of 1.223 °C and one of six out. The pulse and the zone offsets move the plate on purpose.

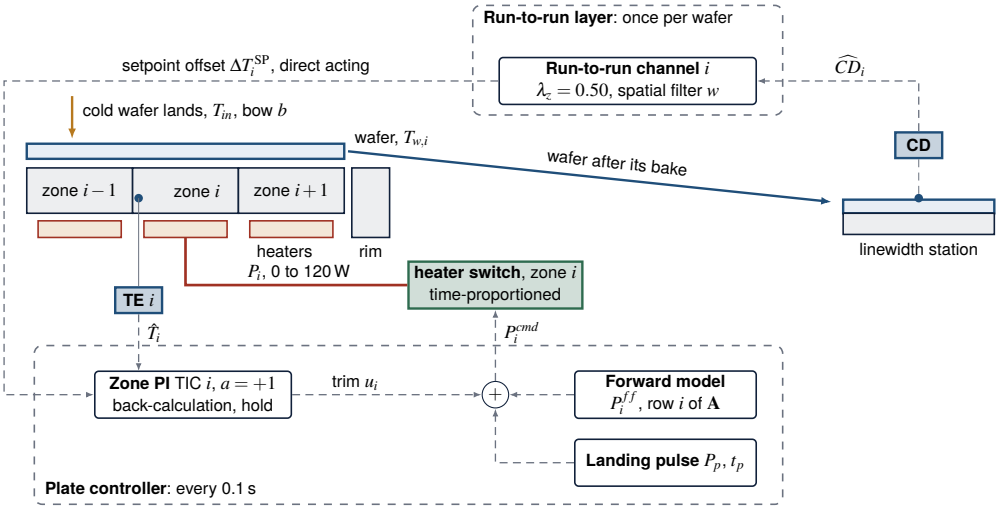


Figure 5.5 One of the hundred zone loops of the simulated plate drawn with its hardware, and the run-to-run loop above it. The thermocouple TE-i in the plate feeds the zone controller TIC-i, a velocity-form PI loop with $a = +1$ every 0.1 s. Its trim is added to the forward-model power and to the timed landing pulse, and the sum drives the heater of zone i. Once per wafer the linewidth map from the measurement station moves the zone setpoint by ΔT_i^{SP}, a direct-acting channel. Notice that no instrument of the inner loop sees the wafer.

5.7 Technology in Practice

The loops of Section 5.5 run on a hundred thermocouples, a hundred heaters, the plate's own controller and, one level up, a linewidth measurement and the run-to-run system that reads it, drawn together in Figure 5.5.

What Measures It

Each zone reads one thermocouple, which the model of Section 5.3 reduces to the plate temperature plus a noise of 0.02 °C in the graded challenge and a fixed bias on zone 0.

> **TECHNOLOGY** **Sensor: A Thermocouple in the Plate**
>
> Two wires of different metals joined at one end produce a voltage at the other end that is roughly proportional to the temperature of the junction. Wherever the wires meet the instrument a second junction forms, so every precision instrument compensates for the temperature of that reference junction. A thermocouple is usually less accurate than a resistance thermometer, but it is rugged and spans a wider range [17]. Its most common failure is to break open, called burnout. An open input picks up electrical noise that reads as a wildly varying temperature, so the instrument is set to drive its reading to the top or the bottom of the scale instead, whichever is safer. On a furnace, where too much heat is the danger, the reading is sent high and the controller then cuts the heat [17].

The model's thermocouples never burn out. They fail only by the bias of scenario 10, the fault that a loop satisfies by holding the plate wrong. The product is measured on every field of every wafer with a repeatability of 0.40 nm and no queue, the most optimistic assumption of the chapter.

What Moves It

Each heater is modeled as a power between zero and 120 W, delivered at once, scaled by the zone's heater health and the supply multiplier of Equation (5.15).

> **TECHNOLOGY** **Final Element: A Heater Switched On and Off**
>
> An electric heater is a final control element, like a control valve or a variable-speed motor. Where a heating element can only be turned fully on or fully off, the controller sends a time-proportioned output instead of a continuous one: a pulse train whose duty cycle, the share of each period spent on, runs from 0 to 100 % [17]. The switching is done without moving parts. In an output channel for alternating current the controller lights a diode inside an optocoupler, which turns on a light-sensitive semiconductor switch, a TRIAC, and power flows to the load [17]. In the heater fault of this chapter the command sits at the full rating while the zone stays cold, and the saturation flag, not the temperature map, names the zone.

The continuous power of the model stands for the average that such a pulse train delivers, and Section 5.3.6 names the switching it leaves out.

Where the Controller Runs

The zone loops run in the plate's own controller, and the run-to-run channels run above it in the fab's advanced process control system. Control loops in operational technology run with cycle times from milliseconds to minutes, many on real-time operating systems, and what counts as real time depends on the application and has to be stated for each [18]. The app executes the zone loops every 0.1 s against a physics step of 0.05 s, and the run-to-run update once per wafer, every 110 s. In the controller the structure is three kinds of block: a PI block per zone with back-calculation and the integral hold, a sum that adds the forward-model power of Equation (5.13) and the timed pulse, and on the upper layer the hundred-channel integrator with its spatial filter.

> **IN PRACTICE** **Run-to-Run Control in a Fab: Filters and Threads**
>
> In a fab, run-to-run control adjusts a recipe lot by lot or wafer by wafer, and most systems combine it with fault detection and classification under the name advanced process control. The exponentially weighted moving average is the most common filter, chosen for its simplicity and robustness. Lots are grouped into control threads by context, such as product, layer and tool, and only lots of one thread update that thread's estimate. The danger is the number of threads: a product that runs rarely leaves its thread without data while the tool drifts [19]. Each of this chapter's hundred channels is one such filter, with a weight of 0.50, on a single thread.

What Is Modernizing It

Two developments bear on this tool. The first measures the wafer's view of the plate. A wireless sensor wafer carries temperature sensors across its face and records the spatial and temporal uniformity of a bake plate through a bake. One such measurement found a laboratory plate uniform near its center with a rapid drop toward its edges, the shape of the uniform-power map of scenario 1. Such checks matter because the international roadmap asked, for the 32 nm node, for a linewidth sensitivity to bake temperature of 1 nm per degree, well below the deep ultraviolet resists of the time [4]. The resist of this chapter has 3.196 nm/K. The second is state estimation that shares information across contexts. Kalman filter methods have been proposed for linewidth control in lithography, and non-threaded estimators attribute a disturbance to the sum of tool, product and layer states with regression or a Kalman filter, at the price of an observability problem in the matrix that assigns runs to contexts [19].

5.8 Limitations

The model is not intended for plate design. The thermal network of (5.1) has no heater layer, no plate body and no proximity gap as separate capacities, so a real plate's step response is higher order than this one and the identification exercise a designer would run does not transfer. The chemistry is a single Arrhenius reaction with no acid diffusion length, so the model reproduces zone-to-zone linewidth steps that a real resist would blur, and it cannot be used to argue about how many zones are worth cutting [3]. Metrology arrives with no queue and every field is measured, which is the most

optimistic assumption in the application and makes the run-to-run loop of (5.16) converge faster than a production loop does.

One limit is worth stating because a reader may look for it and not find it. The lower bound of zero in (5.15) is documented and is never reached in any supplied scenario. Run as shipped for twenty wafers, the smallest power any healthy zone is commanded across all eleven is 28.54 W, in the per-zone run-to-run case, and the 40 W pulse that completes scenario 6 takes it to 26.21 W. Reaching the floor would take a negative offset of 19.9 °C on one zone against its neighbors, or a checkerboard of 11.2 °C, against a bound of 4 °C. The one smaller number in the collection is a delivery rather than a command, the 3.90 W that leaves the damaged heater in the tenth of a second after the fault lands and before its loop winds up. As shipped, the upper bound is reached in scenario 4 only, where the faulted zone is commanded the full 120 W rating for the rest of the run against an unclipped demand that reaches 394.9 W, and delivers 14.40 W through the damaged element. The saturation flag is raised 2.1 s after the fault and stays raised. The same heater failure fired to complete scenario 9 raises it there as well, and it is clear in every other case.

Every disturbance enters either as a commanded step or as a scheduled event, and never as the slow wander a tool shows in production. Perfect repeatability is itself a fiction, because holding the integration step and the scan period constant returns one trajectory on every machine, which no real control system does. Four properties were kept through every simplification, the spatial frequency response, the boundary, the disturbance with a known arrival time and the gap between the plate and the product, and each was kept because a control decision rests on it. Section 5.3.6 carries the condition under which each assumption fails.

Two disturbances are stylized for clarity. A heater that loses 88 % of its authority in one step is not how a cartridge heater usually fails, although it does isolate the difference between the plate signature and the wafer signature of the same fault. A thermocouple bias that appears at once at half a degree is equally abrupt, and it is there so that the moment of onset is unambiguous on the trend.

5.9 Discussion

The spectral view in Equation (5.9) turns a plate specification into arithmetic. Halving k_x by cutting deeper slots lowers $G(0,0)/G(\pi,\pi)$ from 7.86 to 4.43 and buys spatial authority, and it also makes the bare plate less uniform and moves more of the burden onto the control system. Doubling k_x takes the ratio to 14.71, which the application offers as a disturbance so that the trade can be measured rather than asserted. What the calculation shows is that the useful design variable is the ratio of coupling to loss and not the zone count, so a plate with four hundred zones and the same slots resolves a finer correction and delivers it at the same eight-to-one penalty. That conclusion generalizes to any distributed actuator array whose plant is diffusive [7], and Chapter 11 meets it in pressure, where a ragged radial target is out of reach rather than badly tuned.

The model as written also carries control methods the application does not implement. The 201-state balance with hard power limits on every zone is a natural model predictive control problem, and the literature on multizone thermal processing has addressed exactly the computational difficulty that arises when all hundred zones are solved at once [12]. A moving horizon estimator on (5.1) and (5.4) could estimate the contact map m_i of (5.5) from the plate temperatures during the first seconds of a bake, which would convert the wafer bow from an unmeasured disturbance corrected run to run into a measured one corrected within the wafer. An economic formulation could trade the power cost of a rough correction against the yield value of the linewidth range it buys, using the two gains of (5.10) as the exchange rate.

The wider consequence is that linewidth budgets have tightened faster than plates have improved [15], so a share of the uniformity budget at every advanced node is now spent on thermal control rather than on optics. A bake plate is one of the cheapest pieces of a lithography cell and one of the larger contributors to across-wafer linewidth variation, which is why zone counts rose and why

per-zone correction from a measured map became standard [10, 13]. The argument this application makes is that the next increment of performance is not bought with more zones.

OPEN CHALLENGE Randomness Below the Reach of a Heater

During the bake the photoacid diffuses while it catalyzes the deprotection. Photon shot noise, the blur of secondary electrons and that diffusion make pattern formation random at the scale of single features, roughening the lines and sometimes breaking or bridging them. One exposure now prints trillions of features, so for logic chips the failure rate must stay below a few parts per trillion. Moving away from chemically amplified resists may be required, but the metal oxide resists that would replace them raise new questions of stability and reproducibility [20]. Progress would be a resist and bake that keep their speed with fewer random failures. The plate of this chapter sets the dose of a field and nothing finer.

5.10 Conclusion

This chapter studied decentralized control of a coupled actuator array on a hundred-zone post-exposure bake plate. The simulator integrates 201 states, one temperature per zone, one per wafer element and one for the rim, with the resist linewidth taken from an Arrhenius thermal dose referenced to a calibrated bake at a sensitivity of −3.196 nm/K. The steady-state plate is a graph Laplacian plus an ambient term whose gain runs from 2.857 K/W for a uniform power pattern to 0.3636 K/W for a checkerboard, a ratio of 7.86 fixed by the conductance design rather than the zone count.

A gain tuned against a single loop overshoots 18.8 % when all hundred zones move together, because the two spatial gains differ by 4.43. Uniform power leaves the plate 5.437 °C out corner to center, and the forward conductance model holds the flat profile with the feedback switched off entirely.

A cold wafer landing costs 4.015 nm on every field, mostly in dose rather than in the plate dip, and a 26 W pulse for 6 s with the integrators held recovers all but 0.110 nm of it by overshooting the plate 9.3 °C. Bowed wafers print a 3.765 nm radial signature that a hundred run-to-run channels reduce to 0.745 nm while injecting a spatial roughness of 0.254, which a neighbor average takes to 0.189 and 0.595 nm.

A fuller model would add a separate heater layer and proximity gap, an acid diffusion length, a sampled metrology queue and a rim guard heater. The plate as it stands answers the question of what a distributed actuator can be asked to produce.

KEY POINTS

- Every zone shares heat with its neighbors, so the plate answers a change made everywhere far more strongly than a fine pattern, and more zones do not change that.
- A gain tuned on one loop overshoots 18.8 % when all hundred zones move together, because the two spatial gains differ by 4.43.
- A cold wafer landing costs 4.015 nm on every field, mostly in dose, and a timed 26 W pulse with the integrators held recovers all but 0.110 nm of it.
- A bowed wafer prints a 3.765 nm radial signature that no thermocouple sees, and a hundred run-to-run channels on the measured linewidth reduce it to 0.745 nm.

Where else this applies. A spatial bandwidth limit on an actuator array in Chapter 11 and Chapter 9, and a correction learned from late measurements and filtered in space or time in Volume 8, Chapter 10.

References

[1] E. Bristol, “On a new measure of interaction for multivariable process control,” *IEEE Transactions on Automatic Control*, vol. 11, no. 1, pp. 133–134, 1966.

[2] H. Ito and C. G. Willson, “Chemical amplification in the design of dry developing resist materials,” *Polymer Engineering & Science*, vol. 23, no. 18, pp. 1012–1018, 1983.

[3] M. D. Smith and C. A. Mack, "Examination of a simplified reaction-diffusion model for post-exposure bake of chemically amplified resists," in *Advances in Resist Technology and Processing XVIII*, ser. Proceedings of SPIE, vol. 4345. SPIE, 2001, p. 1022.

[4] J. P. Cain, P. Naulleau, and C. J. Spanos, "Critical dimension sensitivity to post-exposure bake temperature variations in EUV photoresists," in *Emerging Lithographic Technologies IX*, ser. Proceedings of SPIE, vol. 5751. SPIE, 2005, p. 1092.

[5] C. D. Schaper, M. M. Moslehi, K. C. Saraswat, and T. Kailath, "Modeling, identification, and control of rapid thermal processing systems," *Journal of The Electrochemical Society*, vol. 141, no. 11, pp. 3200–3209, 1994.

[6] C. D. Schaper, T. Kailath, and Y. J. Lee, "Decentralized control of wafer temperature for multizone rapid thermal processing systems," *IEEE Transactions on Semiconductor Manufacturing*, vol. 12, no. 2, pp. 193–199, 1999.

[7] B. Bamieh, F. Paganini, and M. A. Dahleh, "Distributed control of spatially invariant systems," *IEEE Transactions on Automatic Control*, vol. 47, no. 7, pp. 1091–1107, 2002.

[8] H. Ito, "Chemical amplification resists: Laboratory curiosity to paradigm," *Journal of Photopolymer Science and Technology*, vol. 20, no. 3, pp. 319–332, 2007.

[9] K. El-Awady, C. D. Schaper, and T. Kailath, "Programmable thermal processing module for semiconductor substrates," *IEEE Transactions on Control Systems Technology*, vol. 12, no. 4, pp. 493–509, 2004.

[10] L. Berger, "Global critical dimension uniformity improvement for mask fabrication with negative-tone chemically amplified resists by zone-controlled postexposure bake," *Journal of Micro/Nanolithography, MEMS, and MOEMS*, vol. 3, no. 2, p. 203, 2004.

[11] A. Tay, W. K. Ho, and N. Hu, "An in situ approach to real-time spatial control of steady-state wafer temperature during thermal processing in microlithography," *IEEE Transactions on Semiconductor Manufacturing*, vol. 20, no. 1, pp. 5–12, 2007.

[12] K. V. Ling, W. K. Ho, B. F. Wu, A. Lo, and H. Yan, "Multiplexed MPC for multizone thermal processing in semiconductor manufacturing," *IEEE Transactions on Control Systems Technology*, vol. 18, no. 6, pp. 1371–1380, 2010.

[13] J. Moyne, E. del Castillo, and A. M. Hurwitz, Eds., *Run-to-Run Control in Semiconductor Manufacturing*. Boca Raton, FL: CRC Press, 2018.

[14] Q. Zhang, K. Poolla, and C. J. Spanos, "One step forward from run-to-run critical dimension control: Across-wafer level critical dimension control through lithography and etch process," *Journal of Process Control*, vol. 18, no. 10, pp. 937–945, 2008.

[15] M. van den Brink, "Continued scaling in semiconductor manufacturing enabled by advances in lithography," in *2019 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2019, pp. 1.2.1–1.2.5.

[16] S. Skogestad, "Simple analytic rules for model reduction and PID controller tuning," *Journal of Process Control*, vol. 13, no. 4, pp. 291–309, 2003.

[17] T. R. Kuphaldt, *Lessons in Industrial Instrumentation*, version 2.33 ed. Open textbook, Creative Commons Attribution 4.0 license, ibiblio.org/kuphaldt/socratic/sinst, 2024.

[18] K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, A. Hahn, S. Saravia, A. Sherule, and M. Thompson, "Guide to operational technology (OT) security," National Institute of Standards and Technology, Gaithersburg, MD, Special Publication 800-82r3, 2023.

[19] A. V. Prabhu, "Performance monitoring of run-to-run control systems used in semiconductor manufacturing," Ph.D. dissertation, The University of Texas at Austin, 2008. [Online]. Available: http://hdl.handle.net/2152/17721

[20] D. Kazazis, J. Garcia Santaclara, J. van Schoot, I. Mochi, and Y. Ekinci, "Extreme ultraviolet lithography," *Nature Reviews Methods Primers*, vol. 4, p. 84, 2024.

5.11 Exercises

5.1 ⌕ Concept check · 2 min

Which change to plate BP-301 lowers the ratio $G(0,0)/G(\pi,\pi)$ of its uniform-pattern gain to its checkerboard gain?

(a) Halving the zone-to-zone conductance k_x by cutting deeper slots

(b) Cutting four hundred zones instead of a hundred, with the same slots

(c) Raising the loop gain K_c of all hundred zone loops together

(d) Doubling the heat capacity C_z of every zone of the plate

5.2 Concept check · 1 min

In scenario 7 every thermocouple reads its setpoint, yet the wafers print a 3.765 nm radial range. Why can no zone loop correct it?

(a) The zone loops are too slow to follow the radial signature
(b) The bow changes only the plate-to-wafer contact, which no plate sensor sees
(c) Thermocouple noise is larger than the temperature error at the edge
(d) The edge heaters are already at their power limit during the bake

5.3 Concept check · 2 min

The zone loops stop integrating while the wafer-drop pulse runs. What would happen if they kept integrating through the pulse?

(a) The plate would overshoot further and the line would print narrower
(b) Nothing measurable, because the pulse is shorter than τ_I
(c) The integrators would wind down and sag the plate after the pulse, printing wider
(d) The trims would saturate at the heater rating and trip the flag

5.4 Theory and calculation · 6 min

Use the constants of Table 5.2 and Equation (5.9).

(a) Compute the plate gain for the uniform pattern, the striped pattern $\mathbf{q} = (\pi, 0)$ and the checkerboard $\mathbf{q} = (\pi, \pi)$, to four significant figures. *Checkpoint:* 2.857 K/W, 0.6452 K/W and 0.3636 K/W.

(b) The disturbance Zone-to-zone conduction multiplies k_x. Find the uniform-to-checkerboard gain ratio at 1, 2 and 0.5 times the nominal coupling. *Checkpoint:* 7.857, 14.71 and 4.429.

(c) Find the heater power per zone that holds a 1 °C checkerboard offset and a 1 °C uniform offset, and the ratio of the two. *Checkpoint:* 2.750 W and 0.350 W, a ratio of 7.86.

5.5 Theory and calculation · 7 min

The plate is held flat at 110 °C with no wafer, ambient 22 °C. Use Table 5.2 and the lattice counts of Section 5.3 (69 zones with four neighbors, 18 with three, 13 with two, 44 rim links in all).

(a) Find the rim temperature that sets Equation (5.2) to zero. *Checkpoint:* 98.42 °C.

(b) With Equation (5.13), find the power a four-, a three- and a two-neighbor zone needs, and the total for the plate. How much more is that than the uniform 30.80 W of scenario 1 delivers? *Checkpoint:* 30.800 W, 32.537 W, 34.274 W; 3156.4 W, 76.4 W more than 3080 W.

(c) Find the rim time constant $C_{rim}/(U_{rim} + \sum_i n_i k_r)$ and the zone time constant $C_z/(U_a + 4k_x)$, and their ratio. *Checkpoint:* 31.6 s and 5.16 s, about six.

5.6 Theory and calculation · 7 min

The wafer settles below the plate during a bake, and the linewidth is read against the calibrated dose $\Theta_0 = 0.917514$ of Equation (5.7).

(a) Evaluate Equation (5.8) at 110 °C, and the linewidth worth of a tenth of a degree. *Checkpoint:* −3.196 nm/K and 0.32 nm.

(b) Find the steady gap $U_w(T_i - T_a)/(h_w + U_w)$ between a plate at 110 °C and the wafer on it, to three decimals. *Checkpoint:* 1.483 °C.

(c) A wafer held isothermally at $110 - 1.483$ degrees for the whole bake has a dose equal to its Arrhenius rate. Find its linewidth to 0.01 nm, and the isothermal temperature that prints exactly 62.00 nm. *Checkpoint:* 63.39 nm; 108.95 °C.

5.7 App challenge · 9 min

Load scenario 6, "The feedforward pulse" (default seed), and run it as loaded at the default 2 s/s until the clock passes 110 s, when the first wafer finishes. Export the CSV (its time column is Time (min)). Then press Reset, set Wafer-drop pulse to 14 W, and run to the same time. Repeat once more from Reset with 40 W.

(a) Read the Mean CD and CD range badges after each run, to 0.01 nm. *Checkpoint:* 62.11 nm and 0.27 nm; 65.59 nm and 0.02 nm; 57.07 nm and 0.80 nm.

(b) In the first file, find the highest value of `plate_mean_C` during the bake and the overshoot above setpoint, to 0.01 °C. *Checkpoint:* 119.29 °C, an overshoot of 9.29 °C.

(c) The shipped pulse holds the plate far from flat. Explain why the setting that prints 62 nm is not the one that flattens the plate dip, and what the pulse is sized on. *Checkpoint:* The pulse is sized on dose, because the dip happens while the wafer is cold and costs little dose.

5.8 App challenge · 8 min

Load scenario 3, "Open loop, and the edge" (default seed). Run at 20 s/s until the clock passes 10 min, then Pause.

(a) Read the Plate range badge and convert it to linewidth with Equation (5.8), to 0.1 nm. *Checkpoint:* 5.437 °C, 17.4 nm.

(b) Set Zone control to 100 PI loops, press Run, and let it run 3 min more. Export the CSV. Find the time from the switch (the first row whose `arch` reads `pi`) to the first row with `plate_range_C` below 0.05 °C, to 0.3 s. Read the Edge zone heater value at the end. *Checkpoint:* 44.7 s; about 34.3 W.

(c) Reload the scenario with the reload button beside the scenario list, repeat (b) with Zone control on Model + PI, and explain why the model is not the faster of the two. *Checkpoint:* 52.8 s; the bumpless switch hands the model's edge power to the trims, which release it only as fast as they integrate.

5.9 App challenge · 8 min

Load scenario 8, "Run-to-run, one channel per zone" (default seed), and run it as loaded at 20 s/s until the clock passes 13 min, when seven wafers are done. Export the CSV. Then load scenario 9, "The frequencies the plate cannot produce", and do the same.

(a) Read the CD range badge of each run, and the last value of `corr_checker` (the spatial roughness c of Equation (5.18)) in each file, to ± 0.01. *Checkpoint:* Scenario 8: 0.87 nm and 0.22. Scenario 9: 0.76 nm and 0.16.

(b) Explain why the spatial filter lowers both numbers without slowing the removal of the radial bow signature. *Checkpoint:* The bow signature is smooth and carries little energy at the wavenumbers the filter removes, while the measurement noise is independent from field to field.

5.10 Application area · 3 min

A production version of BP-301 measures linewidth on nine fields per wafer, and each map reaches the run-to-run loop two wafers late ($m = 2$).

(a) What do most of the hundred channels of Equation (5.16) now have to work from, and what does that do to the case for the spatial filter? *Checkpoint:* No measurement of their own: the correction must be spread from nine sites, so a spatial filter or a smooth basis is needed.

(b) With the gain matched, each channel is the newest-reading loop of Equation (1.6) with $K = \lambda_z$. Is $\lambda_z = 0.5$ inside the stability boundary of Equation (1.7) for $m = 2$? Is the challenge's 0.6? *Checkpoint:* $K_{\max} = 0.618$: 0.5 is stable, 0.6 only just.

5.11 Application area · 3 min

In scenario 4 the zone 34 heater keeps 12 % of its authority from $t = 120\,\mathrm{s}$.

(a) Why does the plate map show one cold zone while the linewidth map shows a defect five fields wide? *Checkpoint:* The plate neighbors are regulated; the wafer above them is not, and conducts sideways through k_w.

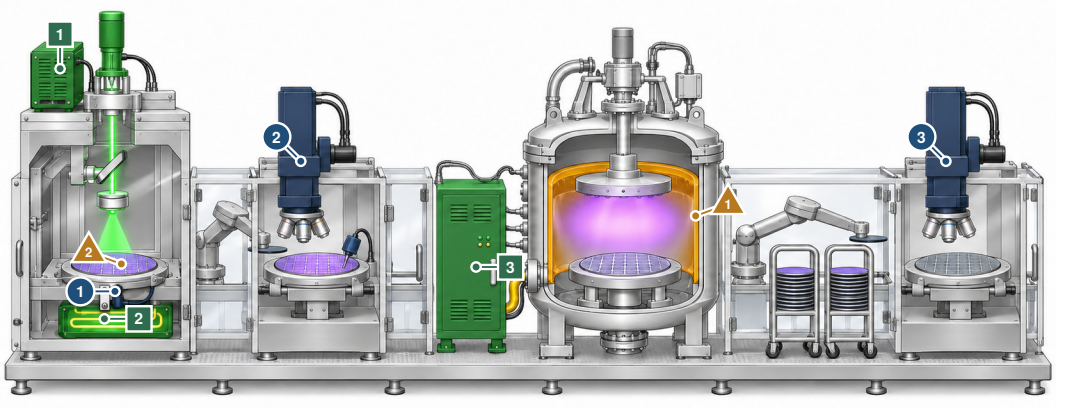
(b) Which logged signal should a maintenance system alarm on to find the heater, and why not the CD range? *Checkpoint:* The saturation flag with the zone command at its rating; the CD map mixes the fault with every other signature.

5.12 Review the agent · 5 min

An AI assistant tunes BP-301 and writes: “A step test on one zone, with its neighbors held by their own loops, gives 0.645 K/W and 5.16 s. SIMC with a closed loop as fast as the plant gives $K_c = 1.55\,\mathrm{W/°C}$ and $\tau_I = 5.16\,\mathrm{s}$. The relative gain is 1.24, so the loops barely interact, and a plate-wide setpoint change will respond like the single-zone test, without overshoot.” Find the flaw, and say how you would check it in the app.

CHAPTER 6

Lithography and Etch Virtual Laboratory for Run-to-Run Control Through a Metrology Queue



The controller moves 1 the exposure dose of the light source 2 the heater under the bake plate 3 the trim time written into the etch recipe

The controller reads 1 the bake plate temperature 2 the microscope that measures the developed lines 3 the microscope that measures the etched lines, two lots late

What disturbs it 1 deposits on the etch chamber walls, emptied by a wet clean 2 a new resist batch that prints the lines wider

Notice the carriers of wafers waiting before the last microscope: the etched linewidth that steers the etch recipe is read only after two more lots have been made. Illustration, not to scale.

IN PLAIN TERMS **Correcting a Recipe With Late News**

The main idea. Chips are made in batches of wafers, and fine lines are printed on each batch with light and then etched. A controller reads how wide the finished lines came out and sets the etching time for the next batch.

The challenge. The answer comes back late, after two more batches have been made. As with a shower whose hot water takes a while to arrive, pushing hard on old news makes the swings grow.

Why it matters. The wait at the measuring tool, and how carefully each result is matched to its own batch, decide how hard the controller may push.

THREE WAYS TO READ THIS CHAPTER

Explore	12 min	The first page, the fact and history boxes of Sections 6.1, 6.2 and 6.4, the try-it boxes of Sections 6.5 and 6.6, the fact box of Section 6.6, and the key points
Study	55 min	The whole chapter with its pause points and the twelve exercises
Practice	15 min	The first page, Section 6.6, Section 6.7, Section 6.8 and the open challenge after it

The control question A lot of wafers is etched every thirty minutes, and its etched linewidth is reported only after two more lots have run. How hard can a recipe controller push on that loop, and how much of the limit comes from the queue and the bookkeeping rather than from the process?

Learning objectives

- Write the EWMA recipe controller as a discrete integral controller and derive its closed loop through a metrology queue of m lots.
- Compute the stability boundary of the loop gain for a given queue and gain ratio, and predict the dead band and decay after a step.
- Explain why an estimator that pairs each reading with the recipe of its own lot removes the queue from the boundary.
- Compare feedforward from the developed linewidth with faster feedback on a measured spread, and price a thinner sampling plan.
- Diagnose two run-to-run loops that share one measurement and fix them by giving each loop its own.

Prerequisites Sections 1.4.3 (run-to-run control), 1.4.4 (metrology delay), 1.4.7 (feedforward between steps), 1.6.2 (velocity form and loop action) and 1.6.4 (scenarios, scorecard and CSV).

The app https://apmonitor.com/apps/litho-etch-r2r/

See also Volume 8, Chapter 9 for a correction that must know which command a late result answers, and Volume 8, Chapter 7 for correction from one part to the next. Volume 5, Chapter 3 for a loop behind a slow analyzer, and Volume 5, Chapter 9 for a sampled loop acting on a valuation a year late.

6.1 Introduction

A control loop is usually characterized by the dynamics of its plant. The loop in this chapter has a plant with no dynamics worth discussing and a measurement that arrives an hour after the material it describes has left the building. A lot of twenty-five wafers is exposed, baked, developed, measured, etched and measured again, and the second measurement is available only after two more lots have passed through the same four stations. Sampling is once per lot and the actuator moves once per lot, so the entire loop lives on a thirty-minute clock with two samples of transport delay in front of it. Nothing about the resist, the plasma or the metrology decides how hard that loop may be pushed. The queue does.

DID YOU KNOW? Half the Steps Are Measurements

In 2018 a review of semiconductor metrology reported that for some products more than 50 % of the manufacturing steps can involve measurement or characterization, and that the share grows as devices shrink and become three-dimensional [1].

The same shape appears wherever the quality measurement is an assay rather than a transmitter. A coating line whose thickness gauge sits two hundred meters downstream, a polymerization reactor whose melt index comes back from a laboratory once a shift, a mineral circuit graded by an off-line particle size analysis, and a casting line inspected after cooling are all plants where the number of units committed between the sample and the correction is the parameter that matters. Textbook treatments of dead time develop the continuous case, where the delay is a fraction of the dominant time constant and the remedy is detuning or a predictor [2]. The discrete case is different because there is no dominant time constant at all. The plant is static within one sample, the controller is a pure discrete integrator, and the delay is an integer count of samples rather than a continuous quantity. What a reader most often carries into this plant, and has to give up, is the belief that a reset rate which is calm on one of these loops will be calm on another. Removing ninety percent of the measured error before the next sample is ordinary practice on a loop whose measurement is live. The same fraction

applied once per lot to the etch recipe of this same line diverges, and the only difference between the two cases is how much material was committed before the answer came back.

Semiconductor recipe control is where that inequality was first written down and where it is still enforced hardest. The exponentially weighted moving average (EWMA) recipe controller entered the industry in the early 1990s [3], its stability and its sensitivity to process gain error were analyzed immediately after [4, 5], and the practical consequences of queued rather than inline metrology have been treated repeatedly since [6, 7]. However, a recursion written on a board as $u_{k+1} = u_k + (\lambda/\hat{b})(r - y_k)$ comes with no plant to try it on, no disturbance that outruns the loop, and no reason to discover that the boundary belongs to the bookkeeping rather than to the process. This chapter makes the delay-limited boundary measurable on an interactive lithography and etch line, on which an estimator that associates each reading with the recipe that produced it moves that boundary by a factor of three, and on which a feedforward path from an upstream measurement removes a variance source that no amount of feedback tuning can filter. The simulator is a six-state disturbance model driving a four-station lot pipeline, with three controllers on three sample rates. Two closed-loop recursions, one per estimator, are each checked against the running simulator to four decimal places in per-lot decay rate, and one plant at one weight of $\lambda_E = 0.95$ is divergent under one estimator and quiet under the other. The EWMA controller and its stability condition are standard, and Sections 1.4.3 and 1.4.4 of Chapter 1 set them out on a smaller line. Section 6.2 traces how recipe feedback replaced the control chart. Section 6.3 gives every equation and constant needed to rebuild the simulator, and Section 6.4 the three controllers with their estimator, feedforward and sampling options. Section 6.5 derives the closed loop through the queue, measures its boundary and shows that the boundary belongs to the estimator. Section 6.6 reports what each scenario measures, from the price of one lot of queue to the graded challenge, Section 6.7 names the instruments and the control layer behind the loops, Section 6.8 states the limitations and Section 6.9 summarizes.

6.2 From Control Charts to Recipe Feedback

Patterning was run open loop for its first two decades. A process engineer qualified a recipe on a chamber, released it, and defended it with a statistical process control chart on the finished linewidth. The chart answered one question, whether the process had changed, and it answered it conservatively by design, because a control chart is built to avoid reacting to common cause variation. Recipes were adjusted at a review, typically once a shift, and the adjustment was an engineering judgment rather than a computation. Every hour of drift between the sample and the review was product.

That arrangement was adequate while the tolerance was a large fraction of the feature. It stopped being adequate when the lithography and etch tools began to drift on a time scale shorter than the review interval. Chamber walls accumulate and lose polymer over hours and change the etch rate and the uniformity with them [8], a resist batch prints a different linewidth at the same dose, and the post-exposure bake plate carries a linewidth sensitivity of several nanometers per degree in a chemically amplified resist [9]. None of those sources is reported by any instrument on the tool that causes them. The only evidence that any of them moved is the linewidth itself.

Run by run control is the answer that was adopted. The controller reads the finished measurement of a lot, updates an estimate of where the process currently sits, and writes the next recipe from that estimate [3]. The estimator chosen was an exponentially weighted moving average, because it is one number of state, it needs no process model beyond a gain, and its weight λ is a single interpretable trade between chasing drift and chasing noise. How the method was proposed, analyzed and collected in its first decade is what the History box tells. A review of automatic control across microelectronics manufacturing placed recipe feedback beside the regulatory layer rather than inside it [10].

Table 6.1 Roles, Nominal Values and Operating Ranges. The Nominal Column Is the Design Recipe From Which Each Scenario Departs.

Symbol	Role	Nominal	Range
CD_{AEI}	CV, once per lot, m late	45.00 nm	38–55 nm SP
CD_{ADI}	CV, once per lot, no delay	62.00 nm	52–74 nm SP
T_{PEB}	CV, continuous	110.00 °C	95–125 °C SP
t_{trim}	MV, once per lot	10.00 s	0–40 s
E	MV, once per lot	28.00 mJ/cm^2	16–44 mJ/cm^2
P_h	MV, continuous	369.6 W	0–1200 W
θ	state, DV, unmeasured	0.5	0–1
δ	state, DV, unmeasured	0 nm	±4 nm/1000 min
g_E	state, DV, unmeasured	1.000	0.92–1.08
F	state, DV, unmeasured	0 µm	0–0.15 µm
β_M	state, DV, unmeasured	0 nm	±1.5 nm
q_{load}	DV, unmeasured	0 W	±40 W

HISTORY From Control Chart to Recipe Feedback (1991 to 2001)

Run-to-run control, a recipe adjusted lot by lot or wafer by wafer, was first proposed for chip fabrication in 1991 [11]. A 1995 framework combined statistical process control with feedback in three parts: a gradual mode for drifts, such as deposits building up in a reactor, a rapid mode for sudden shifts, such as after maintenance, and a control chart that chose between them. On an epitaxy reactor the rapid mode recovered within three runs and the gradual mode cut the variation by a factor of 2.7 [3]. A 1993 analysis gave a simple condition on the weight and the estimated process gain that keeps the controller stable [4]. Work on chemical mechanical polishing reported practical barriers: limits on machine settings, recipes rounded to the steps a tool accepts, and the choice of the weight [12]. A literature review with extensions followed in 1997 [5], and an edited volume collected the field in 2001 [13].

Two refinements matter for this application. The first is the double EWMA, which carries a drift term as well as an intercept so that the loop tracks a ramp without a constant offset, at the cost of a second tuning constant and a transient that is longer than the single EWMA transient [14]. The second is the explicit treatment of metrology delay. Once the measurement of a lot is unavailable until several later lots have run, the loop acquires a pure sample delay, the largest usable weight collapses, and the age of the measurement becomes a tuning input in its own right [6, 7]. Casting the EWMA controller in an internal model control framework makes the same trade explicit as a performance and robustness choice [15]. Integrated metrology, which places the measurement on the process tool and drives the queue toward zero, is the hardware answer to the same inequality, and the alternative software answer is to associate every reading with the tool, chamber and recipe context that produced it so that the controller is working on a matched pair [16]. This application implements the single EWMA with an intercept only, offers both estimators, and exposes the queue length as a control the reader sets.

6.3 Process Model

The line is four stations in series. Litho cell LC-101 coats, exposes at dose E, bakes on plate PEB-101 and develops. Scanning electron microscope CD-SEM-A measures the developed resist linewidth, the after-develop inspection or ADI. Plasma etcher ET-201 trims the resist for t_{trim} seconds and transfers the pattern. CD-SEM-B measures the finished linewidth, the after-etch inspection or AEI. One lot of N_w wafers enters the line every T_{takt} minutes and every station is occupied by a different lot at any instant. Every symbol that appears below is drawn on the line itself, together with the two correction paths, in Figure 6.1.

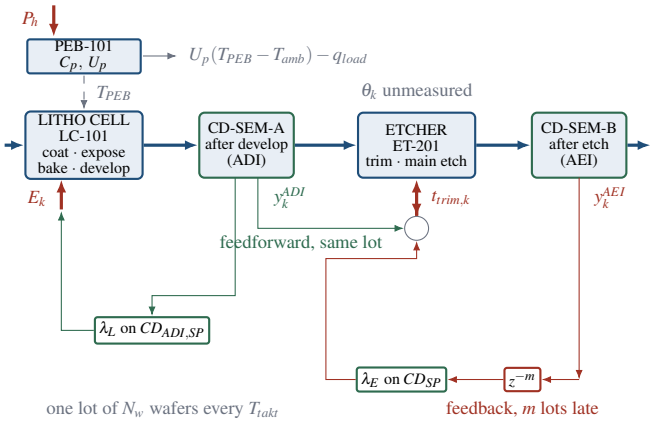


Figure 6.1 The four-station patterning line and its two recipe loops. Lot k is exposed at dose E_k on a plate held at T_{PEB}, measured as y_k^{ADI}, etched for $t_{trim,k}$ in a chamber whose wall condition θ_k is reported by no instrument, and measured again as y_k^{AEI}. The feedback path reaches the etcher only through the queue z^{-m}, while the feedforward path adds $(y_k^{ADI} - CD_{ADI,nom})/\hat{r}$ to the recipe of the same lot one station before that lot is etched.

What the three loops act on and act against is collected in Table 6.1. Every disturbance in the table is unmeasured, which is the defining property of the plant. There is no instrument anywhere on the line that reports the chamber wall condition, the lithography drift, the scanner dose calibration error, the defocus or the metrology offset, and the only evidence that any of them moved is a linewidth arriving late. Time is carried in minutes, linewidth in nanometers, dose in mJ/cm^2, trim time in seconds and plate temperature in degrees Celsius. Six states are integrated continuously and the lot arithmetic is evaluated once per T_{takt}.

🕮 **Pause.** Try Exercise 6.1 before reading on.

6.3.1 The Developed Linewidth

Each lot is a single number. The developed resist linewidth of lot k is the algebraic expression of Equation (6.1).

$$CD_{ADI,k} = CD_{r0} + s_E \ln\left(\frac{g_E E_k}{E_0}\right) - c_F F^2 - k_T\,(T_{PEB} - T_0) + \delta + \varepsilon_{L,k} \tag{6.1}$$

where CD_{r0} is the resist linewidth at nominal dose, best focus and nominal bake, s_E is the exposure latitude in nanometers per unit natural logarithm of dose, g_E is the scanner dose calibration ratio, c_F is the Bossung curvature, F is the defocus, k_T is the bake sensitivity, δ is the accumulated lithography drift and $\varepsilon_{L,k}$ is a zero mean Gaussian lot-to-lot lithography disturbance of standard deviation σ_L. Both sides carry nanometers. Dose enters logarithmically because the resist responds to the logarithm of exposure, which is why exposure latitude is quoted in nanometers per percent of dose rather than per millijoule. At the design point a 1 % dose change moves the developed linewidth by 0.2388 nm and a step from 28.00 to 29.00 mJ/cm^2 moves it by –0.8422 nm.

Focus enters as a square because the Bossung curve through focus has its extremum at best focus [17]. A defocus of either sign moves the printed line in the same direction, so a linear feedback controller can trim the mean back but can never recover the sign of the error and can never say which way to move the stage. A defocus of 0.09 µm narrows the line by 0.9720 nm whether the wafer is high or low. The bake term is linear at 3.2 nm/°C, which puts a tenth of a degree of plate error at 0.32 nm of linewidth and is the reason a bake plate is specified the way it is [9].

6.3.2 The Etcher

The etcher removes a bias from the developed line, set by a fixed component, by the resist trim time and by the condition of the chamber walls, as shown in Equations (6.2) and (6.3).

$$b_k = b_0 + r_{trim}\, t_{trim,k} + \beta\, \theta \tag{6.2}$$

$$CD_{AEI,k} = CD_{ADI,k} - b_k + \varepsilon_{E,k} \tag{6.3}$$

where b_0 is the bias at zero trim, r_{trim} is the resist trim rate in nanometers per second, β is the additional bias contributed by a fully seasoned chamber, θ is the dimensionless wall condition running from 0 for bare aluminum to 1 for fully seasoned, and $\varepsilon_{E,k}$ is a zero mean Gaussian etch disturbance of standard deviation σ_E. The trim time is the manipulated variable of the loop this chapter is about and r_{trim} is its true process gain. At the design recipe of 28.00 mJ/cm^2 and 10.00 s on a chamber at $\theta = 0.5$ with the plate at 110.00 °C, Equation (6.1) returns exactly 62.00 nm, Equation (6.2) returns exactly 17.00 nm and Equation (6.3) returns exactly 45.00 nm. The application verifies those three identities and the vanishing of every state derivative at startup, and the manual scenario holds them: after twenty lots by hand at 369.6 W of plate power the measured plate temperature is 110.000 000 °C, the developed linewidth is 62.000 000 nm and the etched linewidth is 45.000 000 nm.

The wall condition is the only slow state on the etch side and it relaxes toward whatever the chemistry is driving it to, as shown in Equation (6.4).

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = \frac{\theta_\infty - \theta}{\tau_\theta}, \qquad 0 \le \theta \le 1 \tag{6.4}$$

where θ_∞ is the target condition and τ_θ is the seasoning time constant of 600 min, which is twenty lots. A wet clean sets θ to zero and θ_∞ to one in the same instant, which is a step of −2.0 nm in the bias followed by a ten-hour first-order recovery of +4.0 nm. The state is clipped to $[0, 1]$ after every integration step.

6.3.3 The Post-Exposure Bake Plate

The plate is the one part of the line with a live measurement and continuous dynamics, and it is included so that the contrast is available on the same product characteristic. Its energy balance is Equation (6.5).

$$C_p \frac{\mathrm{d}T_{PEB}}{\mathrm{d}t} = 60\,[P_h + q_{load} - U_p\,(T_{PEB} - T_{amb})] \tag{6.5}$$

where C_p is the plate heat capacity in J/K, P_h is the heater power in watts clipped to $[0, 1200\,\mathrm{W}]$, q_{load} is an unmeasured heat load, U_p is the loss coefficient in W/K and T_{amb} is the track ambient. The factor 60 converts watts and joules, which are per second, to the minute clock the rest of the model uses. The open-loop time constant is $C_p/U_p = 61.9\,\mathrm{s} = 1.032\,\mathrm{min}$ and the steady power at 110.00 °C is $U_p(T_{PEB} - T_{amb}) = 369.6\,\mathrm{W}$.

6.3.4 Drift, Calibration and Metrology

The remaining four states carry what moves underneath the line. Equation (6.6) is the lithography drift and Equation (6.7) is the common form of the three calibration states.

$$\frac{\mathrm{d}\delta}{\mathrm{d}t} = \frac{\rho}{1000} \tag{6.6}$$

$$\frac{\mathrm{d}g}{\mathrm{d}t} = \frac{g_{sp} - g}{\tau_{cal}}, \qquad g \in \{g_E,\ F,\ \beta_M\} \tag{6.7}$$

where ρ is the drift rate in nanometers per thousand minutes, g_{sp} is the value the operator or a timed event has requested, and τ_{cal} is a short relaxation time of 3 min that keeps a commanded step continuous enough for a fixed-step integrator. The dose calibration state is clipped to $[0.5, 1.5]$. A

calibration of $g_E = 0.97$, meaning that the scanner delivers three percent less light than it reports, widens the developed line by 0.7310 nm.

Both scanning electron microscopes are modeled as an offset plus white noise on the lot mean, as shown in Equation (6.8).

$$\begin{aligned} y_k^{ADI} &= CD_{ADI,k} + \beta_M + \varepsilon_{M,k} \\ y_k^{AEI} &= CD_{AEI,k} + \beta_M + \varepsilon'_{M,k} \end{aligned} \tag{6.8}$$

where β_M is the tool matching offset and $\varepsilon_{M,k}$ and $\varepsilon'_{M,k}$ are independent zero mean Gaussian draws of standard deviation σ_M. The controller believes Equation (6.8) completely, so an offset goes straight into the silicon. A CD-SEM offset of +0.8 nm applied to a line running the etch loop at $\lambda_E = 0.30$ settles the true etched linewidth at exactly 44.200 nm against a 45.00 nm target, which is 0.8 nm of real material removed to satisfy an instrument error, as Section 1.4.2 of Chapter 1 sets out. Metrology sampling is by lot. Only lots with $k \bmod n_s = 0$ are measured at either station, where n_s is the sampling interval in lots.

A deterministic across-wafer signature is added to the lot mean for the hundred exposure fields of the displayed wafer map, as shown in Equation (6.9).

$$s_{k,i} = f_S\left(u_i^2 - 0.35\right) + f_R\left(\rho_i^2 - 0.42\right) + 1.7\, f_N\, \eta_{k,i} \tag{6.9}$$

where u_i is the normalized across-slit coordinate of field i, ρ_i is its normalized radius, f_S, f_R and f_N are the slit, radial and field-to-field amplitudes, and $\eta_{k,i} \in [-1, 1]$ comes from an integer hash of the lot and field indices so that the same lot always produces the same map. Equation (6.9) is display only. It does not feed back, and the run-to-run controllers see the lot mean, so this application cannot teach spatial recipe control of the kind that corrects the map itself [18].

6.3.5 Parameters, Initialization and Numerics

All twenty-five constants of the plant appear in Table 6.2. Twenty-two are adjustable from the model panel and three never appear in the interface at all.

Every scenario opens at the design point. The states are initialized to $\theta(0) = 0.5$, $T_{PEB}(0) = T_0$, $\delta(0) = 0$, $g_E(0) = 1$, $F(0) = 0$ and $\beta_M(0) = 0$, and the two recipe estimators are seeded from the manual recipe so that a transfer to automatic does not step the line. Equations (6.4) through (6.7) are advanced by a fixed-step fourth-order Runge-Kutta integrator at 0.02 min (1.2 s), the three controllers execute on a 0.04 min (2.4 s) period, and five samples per simulated minute reach the historian. One generator, seeded at 20260915, supplies every random draw in the model. The engine, its clocks and its seeding are those of Section 1.6.1, so a scenario carrying noise replays to the last digit at any playback speed.

6.3.6 Assumptions

The plant is deliberately thin, because everything discarded was discarded to leave the relationship between sampling, delay and gain visible. Each simplification fails somewhere.

Each station is a static map with an additive lot term. Equations (6.1) and (6.3) have no dynamics inside a lot and no memory from one lot to the next except through the six continuous states. Real linewidth response is nonlinear in dose over a wide range, and only the focus term is kept nonlinear here. The assumption fails as soon as the dose is pushed more than about ten percent from E_0.

One number per lot. Within-wafer and within-lot variation enter only through Equation (6.9), which is display only. A production controller also corrects the shape of the across-wafer map [18], and no configuration of this application can show that.

The etcher is one scalar. Equation (6.2) reduces an entire plasma chamber to θ. A chamber wall state genuinely changes radical recombination and therefore both rate and uniformity [8], so the scalar is a legitimate reduction for the lot mean and a poor one for anything spatial.

The lot cadence is fixed and the pipeline is instantaneous. The line never blocks, starves or reorders, every lot takes exactly T_{takt}, and all four stations of a lot are evaluated at one instant rather than as the

Table 6.2 Constants of the Plant. A Hidden Constant Does Not Appear Anywhere in the Interface.

Symbol	Meaning	Value	Unit	Visibility
T_{takt}	lot cadence	30	min	adjustable
N_w	wafers per lot	25	—	hidden
CD_{r0}	resist CD at nominal dose	62	nm	adjustable
E_0	nominal exposure dose	28	mJ/cm^2	adjustable
s_E	exposure latitude	−24	nm per ln E	adjustable
c_F	Bossung curvature	120	nm/µm^2	adjustable
k_T	bake sensitivity	3.2	nm/°C	adjustable
T_0	nominal bake temperature	110	°C	adjustable
C_p	plate heat capacity	260	J/K	adjustable
U_p	plate loss coefficient	4.2	W/K	adjustable
T_{amb}	track ambient	22	°C	hidden
b_0	etch bias at zero trim	10.00	nm	adjustable
r_{trim}	resist trim rate	0.500	nm/s	adjustable
β	bias from a seasoned chamber	4.00	nm	adjustable
τ_θ	seasoning time constant	600	min	adjustable
σ_M	CD-SEM repeatability	0.45	nm	adjustable
σ_L	lot-to-lot litho noise	0.80	nm	adjustable
σ_E	lot-to-lot etch noise	0.35	nm	adjustable
τ_{cal}	calibration relaxation	3	min	hidden
—	specification half-width	2.0	nm	adjustable
f_S	across-slit signature	0.85	nm	adjustable
f_R	radial etch signature	1.10	nm	adjustable
f_N	field-to-field random	0.40	nm	adjustable
—	value of an in-specification lot	180 000	$	adjustable
—	cost of an out-of-specification lot	22 000	$	adjustable

lot walks down the line. The only transport the model carries is the metrology delay m, which the reader sets rather than inherits from dispatch, and that understates how variable the delay is in a real fab.

Metrology noise is white and zero mean within a tool. Tool-to-tool bias is offered as a switchable disturbance in Equation (6.8), but metrology drift is absent, sampling plans that skip wafers rather than lots are absent, and measurement uncertainty that depends on the feature itself is absent.

The estimators carry an intercept and no drift term. A double EWMA would remove the steady offset against a ramp at the cost of a second constant [14]. The offset is left visible on purpose and is quantified in Section 6.5.

The numerics are more reproducible than a fab. Freezing the step and the seed reproduces one trajectory everywhere, which is convenient and false.

6.4 Controllers as Implemented

Three loops run on three clocks. The bake plate is an ordinary proportional-integral controller on a 2.4 s scan. The exposure dose and the etch trim time each move once every 30 min. The lot clock is advanced inside the controller, which fires the whole of a lot at one instant whenever simulated time passes a multiple of T_{takt}.

6.4.1 The EWMA Recipe Controller

Write A_k for the etched linewidth lot k would have had with no trim at all, which by Equations (6.2) and (6.3) is $A_k = CD_{ADI,k} - b_0 - \beta\,\theta$. The measured linewidth is then $y_k = A_k - r_{trim} u_k$ where u_k is the trim time applied to that lot. The controller holds one state, an estimate α of A, and writes the recipe that would put the estimate on target using the trim rate $\hat{r}$ it believes, as shown in Equation (6.10).

$$u_{k+1} = \frac{\alpha_k - CD_{SP}}{\hat{r}} \tag{6.10}$$

The intercept form of the EWMA estimator is Equation (6.11), and it is what the application applies when the estimator option is set to the matched lot.

$$\alpha_k = \lambda_E \left(y_{k-m} + \hat{r}\, u^{FB}_{k-m} \right) + (1 - \lambda_E)\, \alpha_{k-1} \tag{6.11}$$

where m is the after-etch metrology delay in lots and u^{FB} is the feedback part of the recipe. When the estimator option is set to the newest reading the application applies Equation (6.12) instead.

$$\alpha_k = \alpha_{k-1} + \lambda_E \left(y_{k-m} - CD_{SP} \right) \tag{6.12}$$

Substituting Equation (6.10) into Equation (6.12) gives the incremental form of Equation (6.13), which is why a run-to-run controller is a discrete integral controller and why λ_E is a reset time measured in lots. It is Equation (1.3) of Chapter 1 with $\hat{b} = -\hat{r}$, $y^* = CD_{SP}$ and the reading m lots old.

$$u_{k+1} = u_k + \frac{\lambda_E}{\hat{r}} \left(y_{k-m} - CD_{SP} \right) \tag{6.13}$$

There is no proportional term and no derivative term in Equation (6.13). With a 30 min sample and two samples of delay, neither would have anything useful to contribute. Substituting Equation (6.10) into Equation (6.11) at $m = 0$ collapses it to Equation (6.12) for any $\hat{r}$ whatever, so the two estimators are the same controller when there is no queue. With a queue they are not, and Section 6.5 measures the difference.

The recipe is clipped to $0 \le t_{trim} \le 40$ s. Anti-windup is by back-calculation on the estimator rather than on the output. Whenever the clip is active the estimate is recomputed from the trim the etcher can actually deliver, $\alpha_k \leftarrow CD_{SP} + \hat{r}\, u_{clipped}$, so the integral state can never advance past what the actuator will honor. In manual mode both estimators are driven to follow the operator, with $\alpha \leftarrow CD_{SP} + \hat{r}\, t_{trim,manual}$ evaluated every controller scan, which is the discrete equivalent of the tracking that gives the continuous loop its bumpless transfer. In automatic the manual trim and dose stations are set to the applied recipe after every controller execution, so a return to manual is bumpless as well. A change of $\hat{r}$ in automatic rescales the estimate, $\alpha \leftarrow CD_{SP} + (\alpha - CD_{SP})\, \hat{r}_{new} / \hat{r}_{old}$, so the next recipe of Equation (6.10) is the one the old estimate asked for and only the loop gain changes. At the design point a move of $\hat{r}$ from 0.50 to 1.00 nm/s leaves the trim at 10.00 s on every lot. The controller is direct acting on the etched linewidth, because a linewidth above target calls for a longer trim, and the application verifies that sign explicitly. That is action $a = -1$ in the convention of Section 1.6.2. The dose loop below is direct acting for the same reason, since more dose prints a narrower line.

6.4.2 The Dose Loop

The dose loop reads the developed linewidth, which has no queue in front of it, and corrects in the logarithmic domain as shown in Equation (6.14).

$$\ln E_{k+1} = \ln E_k - \frac{\lambda_L}{s_E} \left(y^{ADI}_k - CD_{ADI,SP} \right) \tag{6.14}$$

Because Equation (6.1) depends on $\ln E$ through the same constant s_E that appears in Equation (6.14), the linewidth change produced by one update is exactly $-\lambda_L$ times the measured error. The loop is therefore linear in its own variable rather than only locally linear, one λ_L works over the whole dose range, and the gain ratio is unity by construction. The dose command is clipped to 16–44 mJ/cm^2. A reader may instead point this loop at the etched linewidth, which is offered because the consequence is worth measuring and is reported in Section 6.5.4.

6.4.3 Feedforward From the Developed Linewidth

The developed linewidth of a lot is measured before that lot is etched, so an incoming resist error can be cancelled at the etcher instead of being waited out through a loop that will not see it for m more

lots. The applied recipe is Equation (6.15).

$$t_{trim,k} = u_k^{FB} + \underbrace{\frac{y_k^{ADI} - CD_{ADI,nom}}{\hat{r}}}_{u_k^{FF}} \tag{6.15}$$

where $CD_{ADI,nom}$ is the developed-linewidth setpoint the etcher was set up to trim. The correction exists only for a lot that was measured, so on the lots a sampling plan skips the etcher receives no feedforward at all. Substituting Equation (6.15) into Equations (6.2) and (6.3) with $\hat{r} = r_{trim}$ gives Equation (6.16).

$$\begin{aligned} CD_{AEI,k} = CD_{ADI,nom} - b_0 - r_{trim} u_k^{FB} - \beta\theta \\ - \beta_M - \varepsilon_{M,k} + \varepsilon_{E,k} \end{aligned} \tag{6.16}$$

Every term carrying $CD_{ADI,k}$ has cancelled, including the whole of $\varepsilon_{L,k}$. The lithography disturbance of standard deviation σ_L has been replaced by the ADI metrology disturbance of standard deviation σ_M, so the arrangement is worth having exactly when the measurement is quieter than the process, which is the usual case. It is the full-gain case of Equation (1.11) in Chapter 1.

> **HISTORY Feedforward From Lithography to Etch (1999)**
>
> Feedforward between two process steps adjusts a recipe to compensate for a deviation measured upstream, and it promises lower run-to-run variance, less rework and less scrap. A 1999 study named two problems with it. Noise in the measuring tool can make the variance under feedforward exceed the variance with no control at all, so the recipe adjustment was detuned by the confidence in the sensor. And changing one recipe parameter can worsen others, so the controller chose among recipes already qualified [19]. Its controller adjusted the recipes of a reactive ion etcher to compensate for variations in the lithography linewidth and reduced the variance of the etch process [20].

Equation (6.11) regresses on u^{FB} and deliberately not on the applied trim of Equation (6.15). Regressing on the applied trim would add $y_{k-m}^{ADI} - CD_{ADI,nom}$ to the estimated intercept, which is a lithography deviation the feedforward path has already removed from the product. The feedback loop would then move the next recipe by λ_E times a deviation that no longer appears in any error, and would spend the following lot undoing what the feedforward did.

Pause. Try Exercise 6.3 before reading on.

6.4.4 The Bake Plate Loop

Plate PEB-101 runs the shared velocity-form controller of Section 1.6.2 as a proportional-integral loop on the filtered measurement, with back-calculation anti-windup, written for this loop in Equation (6.17).

$$\begin{aligned} e_k &= T_{SP} - \hat{T}_k, \qquad & s_k &= s_{k-1} + K_{c,T}\,(e_k - e_{k-1}) \\ c_k &= s_k + I_{k-1}, \qquad & P_{h,k} &= \mathrm{clip}(c_k,\ 0,\ 1200\,\mathrm{W}) \\ I_k &= I_{k-1} + \frac{K_{c,T}}{\tau_{I,T}}\, e_k \Delta t_c + \frac{\Delta t_c}{T_t}\left(P_{h,k} - c_k\right) \end{aligned} \tag{6.17}$$

where $\Delta t_c = 0.04\,\mathrm{min}$ is the scan, s_k is the running sum S_k of Equation (1.13), c_k is the power the algorithm demands, which is v_k of Equation (1.14) with no bias, I is the integral state of Equation (1.15) and $T_t = \tau_{I,T}$ is the tracking time of the back-calculation. The integral is seeded with the 369.6 W that balances the plate. While the heater is at a limit only I is corrected, by $\Delta t_c / T_t = 1/25$ of the clipped excess per scan at the default tuning. There is no derivative term and no slew limit. More heater power raises the plate temperature, so the error is setpoint minus measurement and the loop is reverse acting, action $a = +1$ in the convention of Section 1.6.2. The measurement $\hat{T}$ is the plate temperature through a first-order filter of 0.1 min updated by an exact zero-order hold, which is

stable at any period. The default tuning is $K_{c,T} = 15\,\mathrm{W/^{\circ}C}$ and $\tau_{I,T} = 1.00\,\mathrm{min}$ against an open-loop time constant of 1.032 min. Transfer between manual and automatic is bumpless in both directions, by the tracking that Section 1.6.2 describes.

6.5 Stability Through the Queue

6.5.1 The Delay-Limited Boundary

Let $e_k = y_k - CD_{SP}$ and let $\xi = r_{trim}/\hat{r}$ be the ratio of the true trim rate to the one the controller believes, the gain ratio of Section 1.3. Differencing $y_k = A_k - r_{trim}u_k$ with A held constant and substituting Equation (6.13) gives the closed loop of Equation (6.18).

$$e_{k+1} = e_k - \lambda_E\,\xi\,e_{k-m} \tag{6.18}$$

which is a linear difference equation with the characteristic polynomial of Equation (6.19), writing $K = \lambda_E\xi$ for the loop gain.

$$z^{m+1} - z^m + K = 0 \tag{6.19}$$

The pair is Equation (1.6) of Chapter 1 written for this loop. Placing a root on the unit circle, $z = \mathrm{e}^{j\omega}$, gives a phase condition that fixes the crossing frequency at $\omega = \pi/(2m+1)$ with no reference to K and a magnitude condition $2\sin(\omega/2) = K$, so the boundary is that of Section 1.4.4, Equation (6.20).

$$(\lambda_E\xi)_{\max} = 2\sin\frac{\pi}{2\,(2m+1)} \tag{6.20}$$

Equation (6.20) returns 2, 1, 0.618, 0.445 and 0.347 for m = 0 through 4. The value at two lots is the reciprocal of the golden ratio. The crossing frequency also fixes the period of the marginal oscillation at $2\pi/\omega = 2(2m+1)$ lots, which is ten lots at $m = 2$ and is the ringing period a reader sees on the control chart just inside the boundary.

Three consequences are worth separating. The boundary halves between $m = 0$ and $m = 1$ and falls by only a factor of 1.28 between $m = 3$ and $m = 4$, because $2\sin[\pi/(2(2m+1))]$ tends to $\pi/(2m+1)$, so the first lot of queue is the expensive one. The gain ratio ξ and the weight λ_E enter only as a product, so believing a trim rate half the true one doubles the effective gain and can cross the boundary with no tuning change at all. Neither statement mentions the plant, because Equation (6.18) contains no plant dynamics.

Figure 6.2 verifies Equation (6.20) against the running application. The theory curves are the magnitude of the dominant root of Equation (6.19). The markers are measured on the actual etched linewidth after the 1.2 nm resist step of scenario 3, which carries no noise, with λ_E on the 0.01 grid of its slider and m set before the run. The envelope of the error is taken as $(e_k^2 - e_{k-1}e_{k+1})^{1/2}$, which for one oscillatory mode is proportional to $|z|^k$ whatever the phase, and its decay per lot is read between lots 26 and 56, where the trim stays between 10.0 and 14.5 s.

The fifteen measured points agree with the predicted decay to four decimal places in every case, the largest discrepancy being 0.9771 measured against 0.9770 predicted at $m = 4$ and $K = 0.30$. The cusp on each curve is the gain at which the two dominant roots collide and leave the real axis. Below the cusp the recovery is overdamped and slow, because one real root approaches unity as the gain goes to zero, and above it the error rings at the frequency ω.

Equation (6.20) can also be tested at the boundary itself rather than read off a curve. With λ_E set in the simulator to the predicted value, which the slider can only bracket, the envelope measured between lots 50 and 90 decays at exactly 1.0000 per lot at m = 1, 2, 3 and 4. Three percent below the predicted value the same measurement returns 0.9849, 0.9912, 0.9938 and 0.9952, and three percent above it returns 1.0149, 1.0086, 1.0061 and 1.0047, each the dominant root to four decimals. On the slider grid the envelope decays at λ_E = 0.99, 0.61, 0.44 and 0.34 and grows at 1.01, 0.62, 0.45 and 0.35, so the sign of the growth changes where Equation (6.20) says it does and nowhere else. That the weight and the gain ratio enter only as a product is a separate claim and is measured separately.

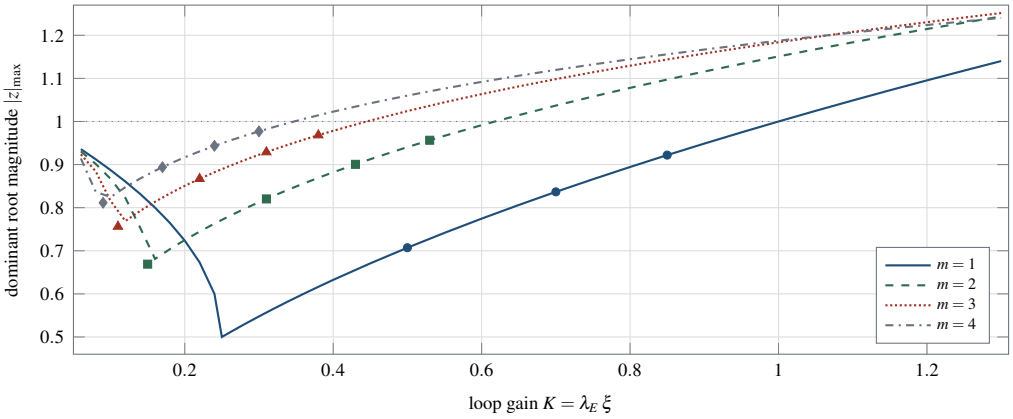


Figure 6.2 Closed-loop decay per lot against loop gain for four queue lengths. Lines are the dominant root magnitude of Equation (6.19), solid for $m = 1$, dashed for $m = 2$, dotted for $m = 3$ and dash-dotted for $m = 4$. Markers are envelope decay rates measured from the application after the resist step of scenario 3, circles for $m = 1$, squares for $m = 2$, triangles for $m = 3$ and diamonds for $m = 4$. The curves cross unity at 1.000, 0.618, 0.445 and 0.347, which are the four values Equation (6.20) predicts for $m = 1$ through 4.

Holding $\lambda_E \xi = 0.30$ at $m = 2$ while moving the believed trim rate through 1.00, 0.50 and 0.25 nm/s, which is ξ = 0.5, 1 and 2, gives a decay of 0.8127 per lot in all three runs, the predicted value. At $\lambda_E \xi = 0.44$, with weights of 0.88, 0.44 and 0.22, the three runs give 0.9066, again the predicted value, and in both sets the three runs agree lot by lot to within 3×10^{-14} nm. Three controllers whose believed trim rate differs by a factor of four recover at the same rate because the product is the same.

6.5.2 Dead Band, Decay and the Residual Against a Ramp

Equation (6.18) also fixes the shape of a step response. A disturbance appearing on lot j is not measured until lot j reaches CD-SEM-B, its reading is not delivered until lot $j + m$, and the recipe it produces is applied to lot $j + m + 1$. Exactly $m + 1$ lots are therefore produced at full error before the recipe moves at all. In the noiseless step scenario a resist batch change widens the developed line by 1.2 nm on lot 6, lots 6, 7 and 8 come out at an error of exactly 1.200 nm, and the feedback trim first moves from 10.000 to 10.720 s in time for lot 9, which is λ_E times the error divided by $\hat{r}$. The errors on lots 9, 10, 11 and 12 are 0.840, 0.480, 0.120 and −0.132 nm, which Equation (6.18) reproduces exactly at $K = 0.30$.

TRY IT IN THE APP 3 · One Step, No Noise

Load the scenario and press Run. Watch the control chart of the etched linewidth after the resist batch changes at 150 min. Lots 6, 7 and 8 come out 1.200 nm wide, because their readings are still in the queue. Then the trim steps from 10.000 to 10.720 s, and the error falls to 0.840, 0.480 and 0.120 nm before it overshoots to −0.132 nm on lot 12.

After the dead band the envelope decays geometrically at the dominant root magnitude, which is 0.8127 per lot at $m = 2$ and $K = 0.30$. Measured successive peaks of the same run give 0.8168, 0.7883, 0.8205, 0.8189, 0.8175 and 0.8162 per lot, spaced six or seven lots apart against the half period of 6.78 lots that the angle of the dominant root predicts. The error falls permanently below ten percent of the step 11 lots after the disturbance, which is 5.5 h of production. A useful contrast is that $(1 - \lambda_E) = 0.70$ is the decay the same controller would achieve with no queue, so two lots of delay slow the recovery by a factor of $\ln 0.70 / \ln 0.8127 = 1.72$ on top of the three lots lost to the dead band.

Against a ramp the intercept-only estimator leaves a constant offset. If the disturbance advances by d nanometers per lot, the recipe must advance at the same rate in steady state, and Equation (6.13) supplies that rate only through $\lambda_E e$, so the residual is Equation (6.21), the result of Equation (1.4) in Chapter 1.

$$e_\infty = \frac{d}{\lambda_E} \tag{6.21}$$

Equation (6.21) contains no m, so the queue costs nothing at all in steady-state ramp following. The drift scenario runs $\rho = 2.5$ nm per 1000 min, which is $d = 0.075$ nm per lot, and the measured residuals at $\lambda_E = 0.15$, 0.30, 0.45 and 0.60 are 0.5000, 0.2500, 0.1667 and 0.1255 nm against the 0.5000, 0.2500, 0.1667 and 0.1250 nm that Equation (6.21) predicts. The last of the four is the one still converging, because at $\lambda_E \xi = 0.60$ and $m = 2$ the dominant root is 0.9914 and the ring decays slowly. Lengthening the window from 50 to 200 lots moves that reading from 0.1260 through 0.1255 to 0.1249 nm, so it is approaching d/λ_E and not settling anywhere else. From lot 203 the ramp holds the trim at its 40 s limit, and from there on the residual stops being a control result at all. A double EWMA removes this term [14] and is not implemented, so the residual is available to be measured.

6.5.3 The Boundary Belongs to the Estimator

Equation (6.20) is a property of Equation (6.12), which corrects a fraction of whatever the newest reading says with no record of which lot produced it. The intercept form of Equation (6.11) regresses each reading against the recipe that actually made that lot, and it gives a different closed loop. Substituting $y_{k-m} = A - r_{trim} u_{k-m}$ into Equation (6.11), eliminating α with Equation (6.10) and writing the result in the error gives Equation (6.22).

$$e_{k+1} = (1 - \lambda_E)\, e_k + \lambda_E\, (1 - \xi)\, e_{k-m} \tag{6.22}$$

The delayed term in Equation (6.22) carries the factor $(1 - \xi)$ rather than the factor 1 of Equation (6.18). When the trim rate is known the delayed term vanishes identically, the loop collapses to the first-order recursion $e_{k+1} = (1 - \lambda_E) e_k$, and it is stable for every $0 < \lambda_E < 2$ at any queue length whatever. The queue still costs $m + 1$ lots of dead band and still costs response time against a drift. It no longer costs stability.

Measurement on the application at $m = 2$ confirms both statements. With $\lambda_E = 0.95$ and the newest-reading estimator the error after the resist step reaches 11.76 nm by lot 38, with the trim driven against its 0 s floor from lot 24, and the run-to-run loop is divergent, while the matched-lot estimator on the identical plant, the identical queue and the identical weight drives the error to numerical zero. Successive lot-to-lot error ratios measured with the matched-lot estimator at $m = 2$ are 0.5000 at $\lambda_E = 0.50$, 0.0500 at 0.95, 0.5000 at 1.50 and 0.9000 at 1.90, each constant from lot to lot and each equal to $|1 - \lambda_E|$ to four decimal places. Those are the ratios of a first-order recursion

with no delayed term in it at all, which is what Equation (6.22) predicts at $\xi = 1$, and they hold at $m = 4$ as well.

Nothing about the plant, the metrology or the queue is different between the two cases. The difference is whether the controller records which recipe produced the reading it just received, which is a database join rather than a control law and is exactly what a production advanced process control system maintains as tool, chamber and recipe context [12, 16]. Volume 8, Chapter 9 meets the same bookkeeping on a vial filler whose checkweigher results arrive while later doses are already committed. The factor between the two boundaries at $m = 2$ is $2/0.618 = 3.24$.

6.5.4 Two Integrators and One Measurement

The dose loop and the etch loop may both be pointed at the etched linewidth. Each is a competent controller and each corrects the whole of the error it sees. The dose update of Equation (6.14) moves the linewidth by exactly $-\lambda_L e$ and the trim update of Equation (6.13) moves it by exactly $-\lambda_E \xi e$, so the two loops superpose into Equation (6.23).

$$e_{k+1} = e_k - (\lambda_E \xi + \lambda_L)\, e_{k-m} \tag{6.23}$$

Two integrators on one measurement are one integrator whose gain is the sum, and Equation (6.20) applies to that sum. After the resist step of scenario 3 the measured envelope decay at $m = 2$ is 0.8822 per lot for $\lambda_E = \lambda_L = 0.20$ and 0.8822 per lot for $\lambda_E = 0.40$ with the dose loop switched off, the dominant root at a gain of 0.40, and the two etched-linewidth sequences agree lot by lot to 10^{-13} nm. At $\lambda_E = \lambda_L = 0.30$ the decay is 0.9914 per lot, just inside the 0.618 boundary, and at $\lambda_E = \lambda_L = 0.40$ the sum of 0.80 is outside it and the loop diverges until the trim reaches its 0 s floor on lot 43 and the dose its 16 mJ/cm^2 floor on lot 53. A configuration in which each loop is individually conservative is therefore unstable as a pair, and no single-loop tuning rule reveals it.

Pause. Try Exercise 6.2 before reading on.

The remedy is not detuning but partitioning. With the dose loop returned to the developed linewidth and the etch loop left on the etched linewidth, each has a measurement the other does not touch. Scenario 9 runs both loops at $\lambda_E = \lambda_L = 0.40$ against a 1.5 nm per 1000 min drift, and over its first 40 lots the shared-measurement configuration gives a spread of 12.012 nm, a capability index of 0.03 and 6 lots in specification. The dose and the trim move together in growing swings about nine lots apart. The trim first reaches zero on lot 14, the dose first reaches its 16 mJ/cm^2 floor on lot 32, and on lot 37 the trim sits at its 40 s ceiling with the etched line at 19.35 nm. Returning the dose loop to the developed linewidth and changing nothing else gives 1.846 nm, a capability index of 0.34 and 31 lots in specification over the same 40. Over five further noise seeds the shared configuration gives 4.57 to 8.98 nm with 13 to 17 lots in specification and the partitioned one 0.90 to 1.46 nm with 33 to 39, so the comparison does not rest on one realization.

The scenario carries the feedforward path switched off, and it has to. Equation (6.16) removes CD_{ADI} from the etched linewidth entirely whenever $\hat{r} = r_{trim}$, and with it the only route the dose loop has to the shared error. With feedforward on, the etched linewidth, the spread, the capability index and the in-specification count are identical to four decimal places whichever measurement the dose loop is pointed at, and the interaction this subsection is about becomes invisible in the controlled variable while remaining plainly visible in the dose.

6.6 Results

The eleven scenarios of Table 6.3 are ordered the way they are taught. Each pins a starting mode, a run-to-run configuration, both weights and the believed trim rate, the noise levels and any timed events, and each asks for a prediction before it is run, in the scenario format of Section 1.6.4. The exported record carries the true linewidth beside the measured one, the feedback half of the trim recipe beside the applied trim, all six continuous states, the loop gain $\lambda_E \xi$, and both stability boundaries side by side, the $2\sin[\pi/(2(2m+1))]$ of Equation (6.20) and the 2 of Equation (6.22), so that the

Table 6.3 The Eleven Scenarios and the Result Each One Isolates

#	Scenario	What it shows
1	The line by hand (MANUAL)	three static gains and three delays
2	On target, nothing moving	a fixed recipe on a plant that is not moving
3	One step, no noise	dead band of $m+1$ lots, geometric decay
4	A slow drift, no run-to-run	ramp offset, Equation (6.21)
5	The chamber clean	a step in an unmeasured state
6	Choosing λ_E at $m=2$	the boundary, and the estimator switch
7	What the queue costs, $m=4$	λ_E fixed, only m moved
8	Lithography noise, $\sigma_L = 1.3$ nm	the λ trade and feedforward
9	Two integrators, one measurement	Equation (6.23)
10	One lot in three measured	sampling, and what it does to feedforward
11	Graded challenge, three upsets	clean, resist step and drift together

estimator comparison can be plotted without switching the controller and Equation (6.18) can be rebuilt in a spreadsheet.

Scenario 1 establishes the architecture before any recipe controller runs. A dose change reaches the developed linewidth and the actual etched linewidth on the very next lot, and the measured etched linewidth two lots after that, when CD-SEM-B reports it. A trim change reaches the actual etched linewidth on the next lot and its measurement two lots later, and never touches the developed linewidth at all, which is the structural zero in the gain matrix and the reason the two loops can be partitioned. The plate is the third input and it is on a different clock. Moving the setpoint 0.5 °C in automatic puts the plate inside 0.1 °C of its new value after 0.32 min and inside 0.01 °C after 0.56 min, and the developed linewidth of the next lot moves by exactly −1.600 nm.

6.6.1 What the Queue and the Weight Buy

Table 6.4 collects the application scorecard for the configurations that matter, each the shipped scenario with the stated change made before the run. Every figure is the on-screen spread over the most recent 40 lots, the capability index against a specification half-width of 2.0 nm, and the fraction of all lots produced that fell inside that specification, read at the stated lot count. Figure 6.3 draws the same scorecard as bars.

The queue block reads across in one direction. Holding λ_E at 0.40 and moving only the metrology delay takes the spread from 1.652 to 3.812 nm and the in-specification fraction from 78 to 47 % between three lots of queue and four. Nothing else changed, and over five further noise seeds four lots of queue gives 2.78 to 3.80 nm against 1.24 to 2.10 nm at three. At $m=3$ the gain of 0.40 is inside the boundary of 0.445 and the loop is merely slow. At $m=4$ it is outside the boundary of 0.347 and the loop diverges until the trim command reaches a limit. One lot of queue, with no change to the controller whatever, is the difference between a line that works and a line that does not, and that pair of numbers is the business case for integrated metrology.

> **TRY IT IN THE APP** **7 · What the Queue Costs**
>
> Press Run and let 60 lots pass through the wet clean with four lots of after-etch queue and a weight of 0.40. The scorecard reads a spread of 3.812 nm and 47 % of lots in specification. Reset, set After-etch metrology delay to 3, and run again: 1.652 nm and 78 %, with nothing else changed.

The gain mismatch block reads the same way with the plant held fixed and only the believed trim rate changed. The three configurations have loop gains of 0.25, 0.50 and 1.00 against the 0.618 boundary, and the spread rises from 1.167 to 7.712 nm across them. A controller commissioned correctly and then left alone while the etcher changes is the practical route into the last row, which is the sensitivity the earliest analyses of the EWMA controller identified [4].

Table 6.4 Measured Scorecard. σ Is the Spread of the Most Recent 40 Lots and the In-Specification Fraction Counts Every Lot Produced. Read at the Lot Count in the First Column.

Lots	Configuration	σ (nm)	C_{pk}	In spec
Lithography noise, $\sigma_L = 1.3$ nm				
60	no run-to-run	1.476	0.45	80 %
60	feedback $\lambda_E = 0.15$	1.599	0.41	77 %
60	feedback $\lambda_E = 0.30$	1.852	0.34	67 %
60	feedback $\lambda_E = 0.60$	2.663	0.24	47 %
60	$\lambda_E = 0.30$ + feedforward	**0.850**	**0.76**	100 %
Wet clean of the etch chamber				
50	no run-to-run	0.782	0.59	100 %
50	etch loop $\lambda_E = 0.40$	0.411	1.43	100 %
Lithography drift, 2.5 nm per 1000 min				
60	no run-to-run	0.877	−0.37	45 %
60	etch loop $\lambda_E = 0.30$	0.001	—	100 %
Choosing the weight at $m = 2$, drift and full noise				
60	$\lambda_E = 0.30$, newest reading	1.265	0.50	82 %
60	$\lambda_E = 0.95$, newest reading	8.002	0.00	15 %
60	$\lambda_E = 0.95$, matched lot	1.548	0.41	82 %
Queue length at $\lambda_E = 0.40$ through a wet clean				
60	$m = 2$	1.497	0.39	80 %
60	$m = 3$	1.652	0.37	78 %
60	$m = 4$	3.812	0.08	47 %
Gain mismatch at $\lambda_E = 0.50$, $m = 2$				
60	$\hat{r} = 1.00$ nm/s, $\xi = 0.5$	1.167	0.53	83 %
60	$\hat{r} = 0.50$ nm/s, $\xi = 1$	1.907	0.34	60 %
60	$\hat{r} = 0.25$ nm/s, $\xi = 2$	7.712	−0.00	13 %
Metrology sampling with feedforward on				
50	every lot measured	1.050	0.61	96 %
50	one lot in two	1.270	0.42	86 %
50	one lot in three	1.092	0.44	88 %
50	every lot, feedforward off	1.525	0.43	82 %

+ DID YOU KNOW? A Wrong Model That Helps

From the simulator: in Table 6.4, at a weight of 0.50 behind a two-lot queue, a controller that believes the trim rate is twice its true value holds a spread of 1.167 nm with 83 % of lots in specification. The controller that knows the true rate holds 1.907 nm and 60 %.

The weight block contains the chapter's sharpest comparison. At $\lambda_E = 0.95$ and $m = 2$ the newest-reading estimator produces a spread of 8.002 nm and 15 % of lots in specification. The matched-lot estimator, with the same weight, the same queue, the same noise and the same drift, produces 1.548 nm and 82 %. That is the in-specification fraction of the properly detuned newest-reading loop at $\lambda_E = 0.30$, with a spread 22 % larger because a weight near one passes most of the metrology noise into the recipe. Over five further noise seeds the newest reading at 0.95, the matched lot at 0.95 and the newest reading at 0.30 keep 20 to 27, 83 to 95 and 90 to 100 % of lots in specification.

The two disturbance blocks in the middle of Table 6.4 are the reason the layer exists at all. A wet clean is a step in a variable no instrument reports, and closing the etch loop at $\lambda_E = 0.40$ takes the spread over 50 lots from 0.782 to 0.411 nm and the capability index from 0.59 to 1.43. A 2.5 nm per 1000 min lithography ramp is worse, because a fixed recipe simply walks out of the band. The

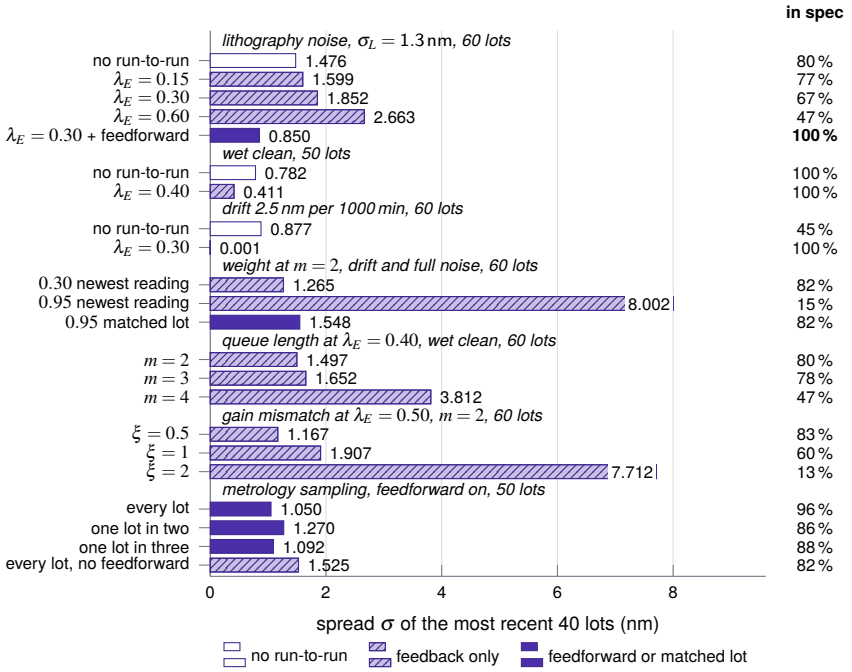


Figure 6.3 The scorecard of Table 6.4 as bars: the spread over the most recent 40 lots, with the share of all lots in specification in the column at the right. Open bars run with no run-to-run loop, hatched bars with feedback alone, solid bars with feedforward or the matched-lot estimator. Notice that the three longest bars, 8.002, 7.712 and 3.812 nm, are the three configurations beyond the delay-limited boundary, by too large a weight, too small a believed trim rate or too long a queue, and that each solid bar is shorter than the hatched bar with the same weight and sampling.

first lot outside specification arrives at $t = 810$ min, against the 800 min the ramp rate and the 2.0 nm tolerance predict, and by 60 lots the capability index is -0.37 with 45 % of lots inside the band. The same loop at $\lambda_E = 0.30$ holds every lot inside it at a spread of 0.001 nm, which makes the capability index meaningless and puts the whole of the error into the 0.250 nm mean offset that Equation (6.21) predicts.

6.6.2 Feedforward Is Not on the Weight Trade Curve

The first block of Table 6.4 is the classical run-to-run trade and then something that is not on it. With 1.3 nm of lot-to-lot lithography scatter the etcher inherits in full, raising λ_E from 0.15 to 0.60 raises the spread from 1.599 to 2.663 nm and drops the in-specification fraction from 77 to 47 %, because a fast discrete integrator writes measurement noise into the recipe and the recipe writes it into the product. The uncontrolled line is better than any of the three feedback tunings, at 1.476 nm and 80 %, which is the honest answer whenever a loop is asked to reject a disturbance that is white at its own sample rate. The margin over $\lambda_E = 0.15$ is small. Over five further noise seeds the uncontrolled spread is below the $\lambda_E = 0.15$ spread in every case but only by 1 to 8 %, inside the scatter of a 40-lot estimate, while $\lambda_E = 0.60$ is the worst of the four in every seed.

Switching feedforward on at $\lambda_E = 0.30$, with the feedback tuning untouched, gives 0.850 nm, a capability index of 0.76 and every one of the 60 lots in specification. Over five further noise seeds the spread is 0.57 to 0.89 nm. The mechanism is Equation (6.16) and not a filter. With λ_E set to zero so that only the feedforward path is active, the measured spread is 0.639 nm against the $\sqrt{\sigma_M^2 + \sigma_E^2} = 0.570$ nm that Equation (6.16) predicts for a 40-lot sample (0.51 to 0.62 nm over

five further seeds), and the entire 1.3 nm lithography term has left the expression rather than being attenuated inside a loop. The remaining difference between 0.639 and 0.850 nm is what the feedback loop adds back by chasing its own after-etch metrology noise.

6.6.3 Sampling Costs Twice

Metrology is expensive and slow, so a production line measures a fraction of its lots. The expected consequence is that the feedback loop corrects once every $n_s T_{takt}$ minutes instead of once every T_{takt}, so the wall-clock recovery from a disturbance lengthens by n_s with no change to λ_E. The unexpected consequence is in Equation (6.15). A lot with no after-develop reading has no feedforward correction available, so $n_s - 1$ lots in every n_s are etched with no knowledge of their own incoming linewidth.

The sampling block of Table 6.4 separates the two effects. Measuring every lot with feedforward on gives 1.050 nm and 96 % in specification. Turning the feedforward path off while still measuring every lot gives 1.525 nm and 82 %, which is the value of the feedforward path alone. Measuring one lot in three with feedforward nominally on gives 1.092 nm and 88 %, which for this seed sits between the two because the path is delivering on a third of the material. The one-in-two row at 1.270 nm is worse than the one-in-three row rather than better, and that ordering should not be read as a result. Each row is one seeded realization of 50 lots, and the standard error of a spread estimated from 40 samples is $\sigma/\sqrt{2 \times 39} = 11\,\%$ of the value, so the two rows are about one and a half standard errors apart. What is outside that band is the 1.525 nm of the last row against the 1.050 nm of the first, although not by a wide margin. Over five further noise seeds feedforward lowers the every-lot spread in four of them, to 0.85 to 1.28 nm from 0.93 to 1.50 nm without it. A sampling plan justified against the feedback loop alone is therefore the wrong calculation, and the error is in the direction of sampling too little.

6.6.4 The Same Integral Action With a Live Measurement

The bake plate is the control in this experiment. An unmeasured −25 W draft across the plate moves its open-loop steady state by $q_{load}/U_p = 5.95\,^{\circ}\mathrm{C}$, which through k_T is 19.0 nm of linewidth, and with the heater held at the 369.6 W that balances the plate the measured temperature falls from 110.0000 to 104.0476 °C over the 50 min that follow. With Equation (6.17) running at the default tuning the same draft produces a peak deviation of 1.199 °C at 0.44 min and the plate is permanently inside 0.1 °C 2.96 min after the draft and inside 0.01 °C after 5.18 min. The peak is 3.84 nm of linewidth equivalent and it lasts three minutes out of a thirty-minute lot.

Compare that with the etch loop meeting the 1.2 nm resist step of scenario 3 at $\lambda_E = 0.30$, which commits three lots at full error and needs 11 lots, or 5.5 h, to fall permanently inside ten percent of the step. Both are integral controllers, both act on the same product characteristic, and the ratio of recovery times is about 110. A wet clean is slower still for the same loop, which needs 26 lots to fall permanently inside ten percent of its 2.0 nm step because the walls go on seasoning underneath it. The entire difference is that one of them is measured every 2.4 s and the other is measured every 30 min with the answer two lots old.

A view of that wet clean recovery in progress appears in Figure 6.4. The disturbance driving all of it shows up in the wall-condition trend strip and in no instrument reading anywhere else, which is the situation run-to-run control was invented for.

6.6.5 The Graded Challenge

Scenario 11 combines a wet clean at 150 min, a resist change at 450 min and a 1.2 nm per 1000 min drift. It is passed when the measured etched linewidth stays within ±1.2 nm of 45 nm from 594 min (about lot 20) to 900 min, with the model parameters and disturbances as given. With the shipped configuration, the etch loop alone at $\lambda_E = 0.08$ without feedforward, the worst measured error from 594 min is 1.89 nm, 0.69 nm outside the band. With the suggested configuration, both recipe loops with $\lambda_E = 0.30$ and $\lambda_L = 0.45$ and feedforward on, it is 0.88 nm, inside with 0.32 nm to spare. Over five further noise seeds the suggested configuration passes every time (0.45 to 0.77 nm) and the

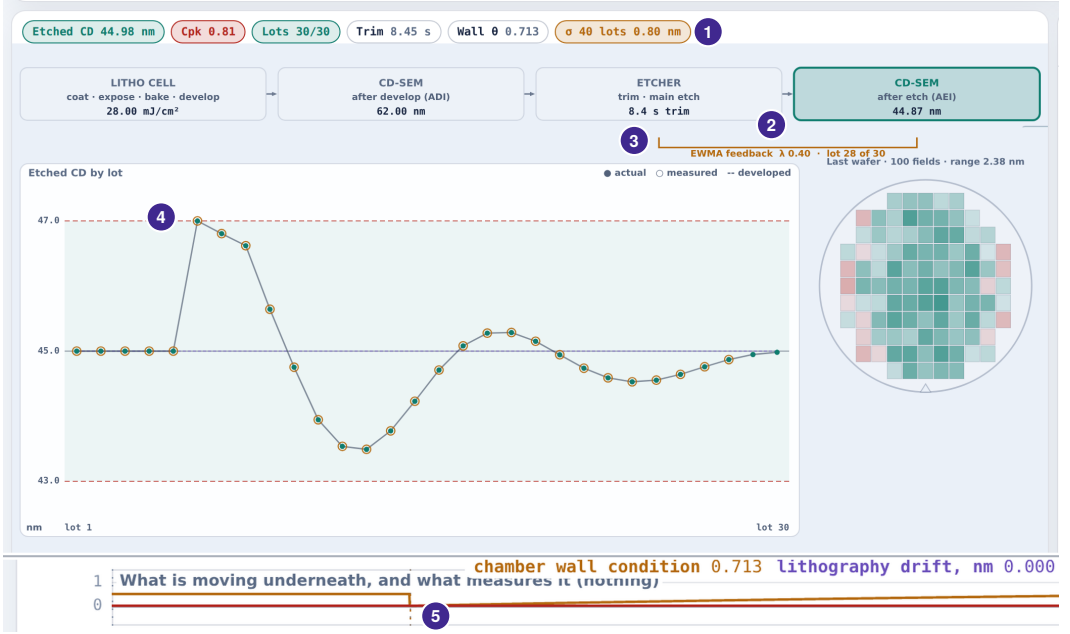


Figure 6.4 Recovery from a wet clean of the etch chamber, scenario 5 with the etch run-to-run loop set to $\lambda_E = 0.40$ before the run and a two-lot metrology queue, at $t = 15.00$ h and lot 30 of 30. The chamber is wet cleaned at $t = 150$ min, the instant lot 6 is etched, so lot 6 is the first lot on bare walls and comes out at exactly 47.00 nm, on the upper specification limit, which the application counts as inside it. The control chart shows the recipe pulling the line back through an undershoot to 43.49 nm at lot 13 before it settles. The trim time has been driven from 10.00 to 8.45 s to follow the wall condition back to 0.713, the reading being acted on is 44.87 nm from lot 28, and the scorecard reads $\sigma = 0.80$ nm with $C_{pk} = 0.81$ and 30 of 30 lots in specification. The view is cropped to the line, the control chart and, below the gray rule, the trend strip of what moves underneath the line. Callout 1 marks the scorecard, 2 the reading CD-SEM-B delivers, 3 the lot that reading belongs to, 4 lot 6, the first lot on bare walls, and 5 the wall condition, which is measured by nothing on the line.

shipped one fails three times. The feedforward path carries most of the improvement, because the resist change is a lithography disturbance that CD-SEM-A reads before the lot is etched, and the etch loop is left with the wet clean, which nothing on the line measures.

6.7 Technology in Practice

The loops of Section 6.4 run on two linewidth microscopes, two recipe numbers, the plate's controller and the fab's run-to-run system, drawn together in Figure 6.5. For a run-to-run loop the sensor is a metrology tool and the final element is a number in a recipe.

What Measures It

Both recipe loops read a critical-dimension scanning electron microscope, CD-SEM-A after develop and CD-SEM-B after etch, which Equation (6.8) reduces to a shared offset β_M and a repeatability of 0.45 nm on the lot mean.

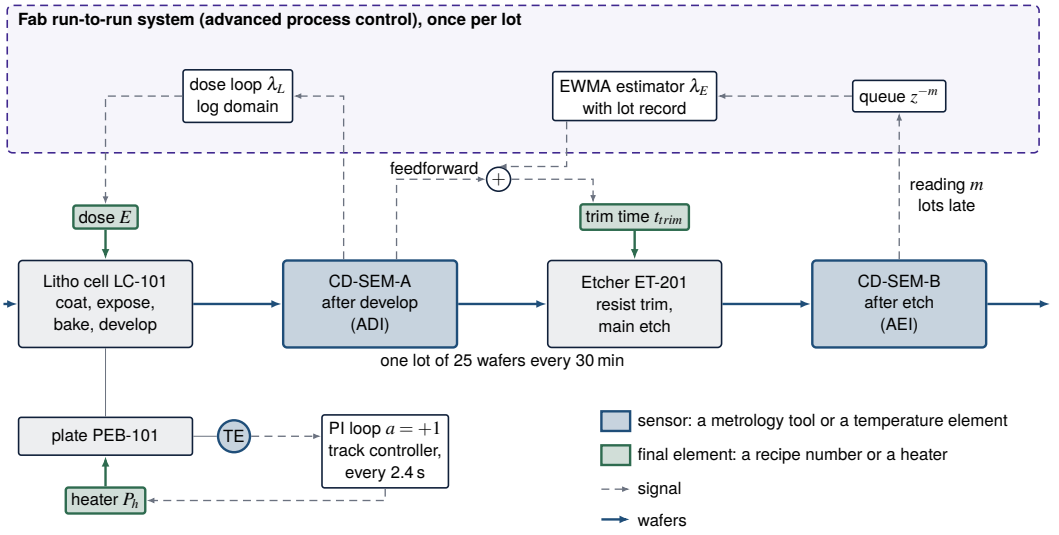


Figure 6.5 The three loops of the simulated line drawn with their hardware. CD-SEM-A reads the developed linewidth of every sampled lot, feeds the dose loop and, through the feedforward sum, the trim of the same lot. CD-SEM-B reads the etched linewidth, and its reading reaches the EWMA estimator only through the queue z^{-m}. Both recipe loops run once per lot in the fab's run-to-run system, while the plate loop runs in the track's own controller every 2.4 s. Notice that the two final elements of the recipe loops are numbers written into a recipe before a lot runs, not devices that move during it.

> **TECHNOLOGY** **Sensor: A Critical-Dimension Scanning Electron Microscope**
>
> A scanning electron microscope sweeps a finely focused electron beam over the wafer, and the secondary and backscattered electrons it releases give the shape and size of a feature with sub-nanometer resolution. Microscopes built for the fab land their electrons at low energy, typically 300 to 800 electronvolts, for repeatable imaging that does not destroy the feature. Their performance is limited less by the electron optics than by drift, vibration, beam damage, charging and contamination. For most work, precision, stability and matching across the fleet of tools count for more than absolute accuracy. Two different techniques can each repeat to better than a tenth of a nanometer and still disagree by more than three nanometers on the same line [1].

The offset β_M stands for the matching error of such a tool, and a reading 0.8 nm high settles the true line at 44.200 nm.

What Moves It

The final element of the etch loop is the trim time of ET-201, clipped to 0–40 s and applied exactly as written. The dose loop writes a second recipe number at the scanner, clipped to 16–44 mJ/cm^2.

> **TECHNOLOGY** **Final Element: A Number in the Etch Recipe**
>
> An etch chamber runs on setpoints for pressures, component temperatures and gas flows. They are developed by experiment for each product and compiled into recipes that are sent to the fabs, where for a long time they were applied open loop [20]. In gate patterning a photoresist trim step narrows the resist line before the main etch, while keeping as much resist thickness as possible for the steps that follow [11]. The same recipe etches differently over time as material deposits on the chamber walls between scheduled maintenance, and the first one or two wafers of a lot often differ from the rest [20].

The wall condition θ of Equation (6.4) is the model's form of that wall deposit, and the wet clean is its maintenance event.

Where the Controller Runs

The plate loop runs in the track's own controller. The recipe loops run above the tools in the fab's advanced process control system, which combines run-to-run control with fault detection and classification [11]. Supervisory and lower-level loops run with cycle times from milliseconds to minutes, and what counts as real time has to be stated for each application [21]. The app executes the plate loop every 2.4 s, while each recipe moves once per lot, every 30 min. In the run-to-run system the structure of this chapter is three blocks: an estimator per loop, a record that pairs each reading with the recipe of its own lot, and a sum that adds the feedforward term of Equation (6.15) to the feedback trim.

> **IN PRACTICE** **Waiting for Metrology, Starting a Thread**
>
> Metrology tools are expensive and shared, so wafers wait to be measured, the most critical first. The delay in front of a run-to-run loop changes from run to run, in most cases between zero and five runs with a mean of one or two, and a longer wait usually raises an alarm for an operator to measure a wafer [11]. A fab runs a separate moving-average controller for each thread, one combination of product and prior processing path. A new thread's first estimate is set by hand or by a pilot run, which costs, or set automatically at the risk of outliers in its first runs [16].

What Is Modernizing It

Three developments bear on this loop. Integrated metrology adds a measuring tool to each process tool so that product can be measured during or after the step, and by 2011 its wide adoption had been held back by set-up cost, limited cooperation between chip makers and tool makers, and longer cycle times [20]. Scatterometry, a variant of ellipsometry, fits the light scattered from an array of lines to a library of simulated results and returns the linewidth and height without an image [1]. Virtual metrology predicts post-process metrology variables from process and wafer state information [1], such as the light emitted by an etch plasma, for every wafer, while real metrology covers only some wafers and arrives hours or even days late [20].

6.8 Limitations

The application is not a lithography simulator. Equation (6.1) is four terms, and a real linewidth model resolves the aerial image, the resist chemistry, the development kinetics and the mask through a physical simulation, so no question about illumination, pitch or mask bias can be asked here. The etcher of Equation (6.2) is one scalar, and questions about the discharge, the neutral chemistry or the profile belong to the plasma etch chamber of Chapter 7. The across-wafer signature of Equation (6.9) is decorative. It does not respond to the controller, so spatial recipe control of the kind that corrects the map itself [18] is outside reach. The estimators carry an intercept and no drift term, so the residual of Equation (6.21) is larger than a production controller would show. Both estimators also assume that the gain ratio is constant and known to whatever accuracy $\hat{r}$ provides, while a production system re-identifies the gain as a second and slower loop that can argue with the first one. The lot cadence is fixed, the line never blocks or reorders, and every lot is one number, so nothing about dispatch, lot splitting or wafer-level sampling can be studied. The economics are a flat value per lot in specification and a flat cost per lot outside it, which is not how a rework decision is actually made. Reproducibility itself is unrealistic, because freezing the seed and the step gives one trajectory on every machine while a fab gives a distribution. Section 6.3.6 names the failure mode of each assumption separately.

Three elements of the scenarios are stylized on purpose. A wet clean that empties the chamber wall state instantly idealizes a maintenance event that takes hours, but it isolates a step in a variable no instrument reports. A metrology queue set with a slider would in practice be an outcome of

dispatch and tool availability, although making it a setting is what allows the boundary to be measured against Equation (6.20) one lot at a time. A believed trim rate set to half the true one is unlikely in a commissioned line, while it is the cleanest available way to separate λ_E from ξ in Equation (6.18).

OPEN CHALLENGE Measuring Devices That Have Gone Three-Dimensional

As transistors shrink and turn three-dimensional, the industry approaches the point where the position and type of each atom in a device need to be known, for billions of devices on each chip. No single instrument has the resolution, speed and low uncertainty to measure every parameter, and different techniques can disagree on the same feature, because their error sources and their definitions of what is measured differ [1]. Progress would be measurements traceable to a common reference, with agreed definitions of each quantity [1]. The loops of this chapter trust one linewidth number per lot completely.

6.9 Conclusion

This chapter studied control on a measurement that arrives late, on a four-station lithography and etch line. The simulator integrates six continuous states covering the etch chamber wall condition, the bake plate temperature, a lithography drift and three calibration errors, and evaluates the lot arithmetic once every 30 min at a design recipe of 28.00 mJ/cm^2 and 10.00 s that puts the developed line at 62.00 nm, the etch bias at 17.00 nm and the etched line at 45.00 nm.

The exponentially weighted moving average recipe controller is a discrete integral controller whose closed loop is $e_{k+1} = e_k - \lambda_E \xi e_{k-m}$, and the stability boundary $2\sin[\pi/(2(2m+1))]$ that follows returns 2, 1, 0.618, 0.445 and 0.347 for zero through four lots of queue. Fifteen envelope decay rates measured on the running application agree with the dominant root of that recursion to four decimal places. Regressing each reading against the recipe that produced that lot replaces the recursion with $e_{k+1} = (1-\lambda_E)e_k + \lambda_E(1-\xi)e_{k-m}$, which with the trim rate known is stable for any weight below 2 at any queue length, and the same plant that diverges at $\lambda_E = 0.95$ with the newest-reading estimator holds 1.548 nm of spread with the matched-lot one.

Feedforward from the developed linewidth removes the lithography term from the etched linewidth outright rather than filtering it, taking the spread from 1.852 to 0.850 nm and the in-specification fraction from 67 to 100 % with the feedback tuning untouched. Two integrators pointed at one measurement add their weights, so a pair that is individually conservative at 0.40 each is unstable at a sum of 0.80 and takes the spread over 40 lots from 1.846 to 12.012 nm as the recipes swing between their limits.

A fuller model would add a drift term to the estimator, a gain re-identification loop, a dispatch model that makes the queue an outcome rather than a setting, and across-wafer control of the signature that is presently display only. On this line the cost of a queued measurement is something the reader measures rather than accepts.

KEY POINTS

- A lot is measured after it leaves the etcher and the answer arrives two lots later, so three lots are made at full error before the recipe moves.
- The largest usable loop gain, $2\sin[\pi/(2(2m+1))]$, is 1, 0.618, 0.445 and 0.347 for one to four lots of queue.
- Pairing each reading with the recipe of its own lot keeps the loop stable for any weight below 2 at any queue length when the trim rate is known.
- Feedforward from the developed linewidth removes the lithography term, taking the spread from 1.852 to 0.850 nm.
- Two integrators on one measurement add their weights: 0.40 each is unstable as a pair.

Where else this applies. Late results matched to their commands in Volume 8, Chapter 9, a slow analyzer in Volume 5, Chapter 3, and run-to-run channels in Chapter 5.

References

[1] N. G. Orji, M. Badaroglu, B. M. Barnes, C. Beitia, B. D. Bunday, U. Celano, R. J. Kline, M. Neisser, Y. Obeng, and A. E. Vladar, "Metrology for the next generation of semiconductor devices," *Nature Electronics*, vol. 1, no. 10, pp. 532–547, 2018.

[2] D. E. Seborg, T. F. Edgar, D. A. Mellichamp, and F. J. Doyle, III, *Process Dynamics and Control*, 4th ed. Hoboken, NJ: John Wiley & Sons, 2016.

[3] E. Sachs, A. Hu, and A. Ingolfsson, "Run by run process control: combining SPC and feedback control," *IEEE Transactions on Semiconductor Manufacturing*, vol. 8, no. 1, pp. 26–43, 1995.

[4] A. Ingolfsson and E. Sachs, "Stability and sensitivity of an EWMA controller," *Journal of Quality Technology*, vol. 25, no. 4, pp. 271–287, 1993.

[5] E. Del Castillo and A. M. Hurwitz, "Run-to-run process control: Literature review and extensions," *Journal of Quality Technology*, vol. 29, no. 2, pp. 184–196, 1997.

[6] R. P. Good and S. J. Qin, "On the stability of MIMO EWMA run-to-run controllers with metrology delay," *IEEE Transactions on Semiconductor Manufacturing*, vol. 19, no. 1, pp. 78–86, 2006.

[7] A. Chen and R.-S. Guo, "Age-based double EWMA controller and its application to CMP processes," *IEEE Transactions on Semiconductor Manufacturing*, vol. 14, no. 1, pp. 11–19, 2001.

[8] T. W. Kim and E. S. Aydil, "Effects of chamber wall conditions on Cl concentration and Si etch rate uniformity in plasma etching reactors," *Journal of The Electrochemical Society*, vol. 150, no. 7, p. G418, 2003.

[9] J. P. Cain, P. Naulleau, and C. J. Spanos, "Critical dimension sensitivity to post-exposure bake temperature variations in EUV photoresists," in *Emerging Lithographic Technologies IX*, ser. Proceedings of SPIE, vol. 5751. SPIE, 2005, p. 1092.

[10] T. F. Edgar, S. W. Butler, W. J. Campbell, C. Pfeiffer, C. Bode, S. B. Hwang, K. S. Balakrishnan, and J. Hahn, "Automatic control in microelectronics manufacturing: Practices, challenges, and possibilities," *Automatica*, vol. 36, no. 11, pp. 1567–1603, 2000.

[11] A. V. Prabhu, "Performance monitoring of run-to-run control systems used in semiconductor manufacturing," Ph.D. dissertation, The University of Texas at Austin, 2008. [Online]. Available: http://hdl.handle.net/2152/17721

[12] D. Boning, W. Moyne, T. Smith, J. Moyne, and A. Hurwitz, "Practical issues in run by run process control," in *Proceedings of SEMI Advanced Semiconductor Manufacturing Conference and Workshop*, 1995, pp. 201–208.

[13] J. Moyne, E. Del Castillo, and A. M. Hurwitz, Eds., *Run-to-Run Control in Semiconductor Manufacturing*. Boca Raton, FL: CRC Press, 2001.

[14] E. Del Castillo, "Long run and transient analysis of a double EWMA feedback controller," *IIE Transactions*, vol. 31, no. 12, pp. 1157–1169, 1999.

[15] S. Adivikolanu and E. Zafiriou, "Extensions and performance/robustness tradeoffs of the EWMA run-to-run controller by using the internal model control structure," *IEEE Transactions on Electronics Packaging Manufacturing*, vol. 23, no. 1, pp. 56–68, 2000.

[16] F. Harirchi, T. L. Vincent, A. Subramanian, K. Poolla, and B. Stirton, "On the initialization of threaded run-to-run control of semiconductor manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 27, no. 4, pp. 515–522, 2014.

[17] J. W. Bossung, "Projection printing characterization," in *Developments in Semiconductor Microlithography II*, ser. Proceedings of SPIE, vol. 0100. SPIE, 1977, pp. 80–85.

[18] Q. Zhang, K. Poolla, and C. J. Spanos, "One step forward from run-to-run critical dimension control: Across-wafer level critical dimension control through lithography and etch process," *Journal of Process Control*, vol. 18, no. 10, pp. 937–945, 2008.

[19] S. Ruegsegger, A. Wagner, J. S. Freudenberg, and D. S. Grimard, "Feedforward control for reduced run-to-run variation in microelectronics manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 12, no. 4, pp. 493–502, 1999.

[20] S. Lynn, "Virtual metrology for plasma etch processes," Ph.D. dissertation, National University of Ireland Maynooth, 2011. [Online]. Available: https://mural.maynoothuniversity.ie/id/eprint/2657/

[21] K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, A. Hahn, S. Saravia, A. Sherule, and M. Thompson, "Guide to operational technology (OT) security," National Institute of Standards and Technology, Gaithersburg, MD, Special Publication 800-82r3, 2023.

6.10 Exercises

6.1 Concept check · 1 min

Which of these quantities moves the etched linewidth of the patterning line but is reported by no instrument anywhere on it?

(a) The chamber wall condition θ of etcher ET-201
(b) The developed linewidth read by CD-SEM-A
(c) The plate temperature of bake plate PEB-101
(d) The etched linewidth read by CD-SEM-B

6.2 Concept check · 2 min

In scenario 9 the dose loop and the etch loop each run at a weight of 0.40, and both are pointed at the etched linewidth behind the two-lot queue. Each loop alone would be stable. Why does the pair diverge?

(a) The dose loop has no queue, so it outruns the etch loop and drags it along
(b) Both correct the same delayed error, so they act as one integrator of gain 0.80
(c) The dose stops at its 16 mJ/cm^2 floor and its integrator winds up there
(d) The feedforward path adds the dose correction to the trim a second time

6.3 Concept check · 2 min

With 1.3 nm of lot-to-lot lithography scatter, which element of the line removes that scatter from the etched linewidth instead of filtering it?

(a) Raising λ_E so the etch loop corrects each lot faster
(b) The PI loop that holds the bake plate at its setpoint
(c) Feeding the measured developed linewidth forward to the trim of the same lot
(d) Pointing the dose loop at the etched linewidth as well

6.4 Theory and calculation · 6 min

Use Equations (6.18) and (6.20) for the newest-reading estimator.

(a) Compute the boundary $(\lambda_E \xi)_{\max}$ for $m = 0$ to 4 to three decimals, and the ratio of the $m = 0$ value to the $m = 1$ value and of the $m = 3$ value to the $m = 4$ value, to two decimals. *Checkpoint:* 2.000, 1.000, 0.618, 0.445 and 0.347; ratios 2.00 and 1.28.

(b) The controller believes $\hat{r} = 0.25\,\text{nm/s}$ while the true trim rate is 0.50 nm/s, and $m = 2$. Find the largest λ_E on the slider's grid of 0.01 that keeps the loop stable. *Checkpoint:* $\xi = 2$, so $\lambda_E < 0.309$: 0.30.

(c) Find the period of the marginal oscillation at $m = 2$, in lots and in hours at a lot cadence of 30 min. *Checkpoint:* 10 lots, 5 h.

6.5 Theory and calculation · 7 min

Scenario 3 has no noise. A resist change widens the developed line by 1.2 nm from lot 6, the queue is $m = 2$, $\lambda_E = 0.30$, $\hat{r} = r_{trim} = 0.50\,\text{nm/s}$, and the feedback trim is 10.00 s before the step.

(a) With the newest-reading estimator, use Equation (6.18) to find the error on lots 6 to 12, to three decimals, and use Equation (6.13) for the trim applied to lot 9. *Checkpoint:* 1.200, 1.200, 1.200, 0.840, 0.480, 0.120 and −0.132 nm; 10.720 s.

(b) With the matched-lot estimator, use Equation (6.22) to find the error on lots 9, 10 and 11, to three decimals. *Checkpoint:* 0.840, 0.588 and 0.412 nm.

(c) Explain why both estimators give 0.840 on lot 9 and part afterward. *Checkpoint:* The first correction answers the first reading in both. From lot 10 the newest reading keeps correcting errors that earlier corrections have already answered.

6.6 Theory and calculation · 6 min

Scenario 4 drifts the lithography at 2.5 nm per 1000 min. The lot cadence is 30 min and lot k is etched at $30(k-1)$ minutes, the tolerance is ± 2.0 nm, the trim rate is 0.50 nm/s, and the line starts at the design point with 10.00 s of trim and the newest-reading estimator.

(a) With a fixed recipe, which lot is the first outside specification, and at what time is it etched? *Checkpoint:* Lot 28, at 810 min.

(b) With the etch loop on, find the settled offset at λ_E = 0.15, 0.30 and 0.45, to three decimals. *Checkpoint:* 0.500, 0.250 and 0.167 nm.

(c) Once the offset has settled the trim must rise at a steady rate. Find that rate in seconds per lot, and the first lot at which $\lambda_E = 0.60$ asks for the 40 s ceiling. *Checkpoint:* 0.15 s per lot; lot 203.

6.7 App challenge · 8 min

Load scenario 6, "Choosing λ, and the boundary the queue sets" (default seed). It runs the etch loop at $\lambda_E = 0.95$ behind a two-lot queue, with full noise and a 1.5 nm per 1000 min drift. Run it as loaded at 20 min/s and press Pause while the Lots run tile reads 60. Lot 60 runs from 1770 to 1800 min, and the tiles hold their values for the whole lot. Then press Reset, set EWMA weight · etch loop to 0.30, and run to lot 60 again. Last, press Reset, set the weight back to 0.95, set The etch estimator uses to The matched lot (in Advanced it sits under Metrology and sampling), and run to lot 60 a third time.

(a) Read the σ (40 lots) and Lots in spec tiles after each run, to 0.01 nm and 1 %. *Checkpoint:* 8.00 nm and 15 %; 1.27 nm and 82 %; 1.55 nm and 82 %.

(b) Give the loop gain $\lambda_E \xi$ of each run and the boundary of its estimator, and say which run is outside it. *Checkpoint:* 0.95 against 0.618, outside; 0.30 against 0.618; 0.95 against 2.

(c) The third run matches the second in lots in specification but not in spread. Explain the larger spread. *Checkpoint:* A weight near one passes most of the after-etch metrology noise into the recipe.

6.8 App challenge · 9 min

Load scenario 9, "Two integrators, one measurement" (default seed). Both recipe loops run at a weight of 0.40, feedforward is off, and The dose loop controls (in Advanced) is set to Etched CD. Run at 20 min/s, press Pause while the Lots run tile reads 40, and export the CSV.

(a) Read the σ (40 lots), Cpk and Lots in spec tiles. *Checkpoint:* 12.01 nm, 0.03 and 15 %.

(b) In the CSV, find the first `lot_no` whose `trim_s` is 0 and the first whose `dose_mJ` is 16. *Checkpoint:* Lots 14 and 32.

(c) Press Reset, set The dose loop controls to Developed CD, change nothing else, and run to lot 40 again. Read the same three tiles and explain the change. *Checkpoint:* 1.85 nm, 0.34 and 78 %. Each loop now has its own measurement.

6.9 App challenge · 8 min

Load scenario 8, "Lithography noise, and what feedforward does to it" (default seed). It has $\sigma_L = 1.3$ nm and the etch loop at $\lambda_E = 0.30$. Run at 20 min/s, press Pause while the Lots run tile reads 60, and read the σ (40 lots) and Lots in spec tiles. Press Reset, set Feedforward ADI → etch to On, and repeat. Press Reset again, keep feedforward On, set EWMA weight · etch loop to 0, and repeat.

(a) Give σ and Lots in spec for the three runs, to 0.01 nm and 1 %. *Checkpoint:* 1.85 nm, 67 %; 0.85 nm, 100 %; 0.64 nm, 100 %.

(b) Compute the spread Equation (6.16) predicts with feedforward alone, to three decimals, and judge the third run against it, given that a spread estimated from 40 lots has a standard error of about 11 %. *Checkpoint:* $\sqrt{\sigma_M^2 + \sigma_E^2} = 0.570\,\text{nm}$, and 0.64 is about 12 % high, about one standard error.

(c) Explain why the second run is wider than the third. *Checkpoint:* The feedback loop chases its own after-etch metrology noise.

6.10 Application area · 3 min

Table 6.4 runs the etch loop at $\lambda_E = 0.40$ through a wet clean. With three lots of queue the line holds a spread of 1.652 nm and 78 % of lots in specification. With four it holds 3.812 nm and 47 %.

(a) Explain from Equation (6.20) why one lot of queue makes this difference. *Checkpoint:* 0.40 is inside 0.445 at $m = 3$ and outside 0.347 at $m = 4$.

(b) Name two changes that would let the line run a weight of 0.40 with confidence, other than lowering the weight. *Checkpoint:* Shorten the queue (integrated metrology), or use the matched-lot estimator.

6.11 Application area · 3 min

A metrology group proposes measuring one lot in three on this line to free CD-SEM time, and supports it with the feedback loop alone: the loop will simply correct less often.

(a) Name the two costs of the plan on this line. *Checkpoint:* The feedback loop corrects once every 90 min, and two lots in three get no feedforward.

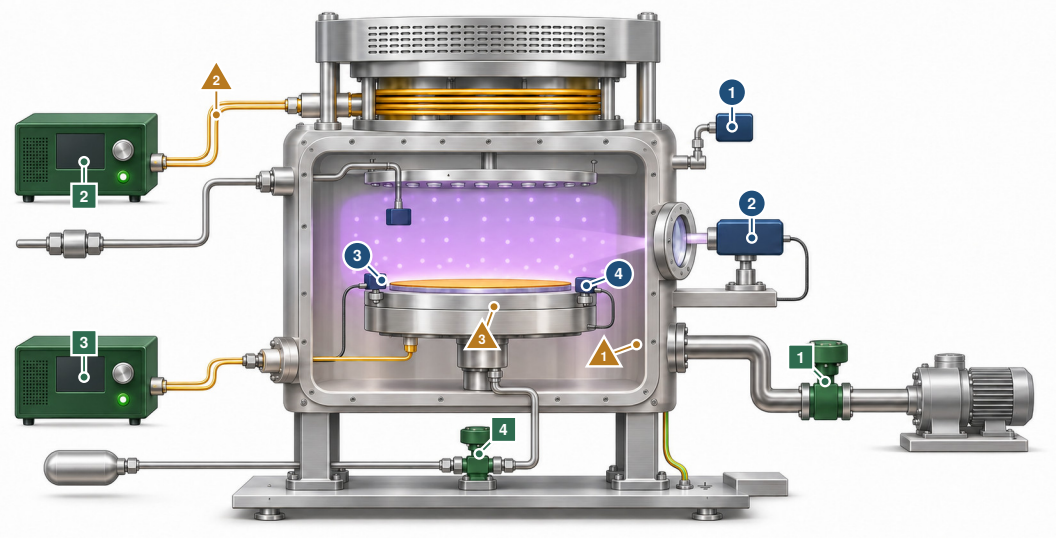
(b) Which of the two measurements would you keep on every lot, and why? *Checkpoint:* The developed linewidth, because it feeds forward to the same lot with no queue.

6.12 Review the agent · 5 min

An AI assistant looks at scenario 4 and writes: “The newest-reading loop leaves $d/\lambda_E = 0.25\,\text{nm}$ against the drift at $\lambda_E = 0.30$, and it cannot go above 0.618 at $m = 2$. Switch The etch estimator uses to The matched lot, which is stable up to $\lambda_E = 2$ at any queue when the trim rate is known, and set $\lambda_E = 0.75$. The residual then falls to $d/\lambda_E = 0.10\,\text{nm}$.” Find the flaw, and say how you would check it in the app.

CHAPTER 7

A Plasma Etch Simulator for Recipe Trades and Endpoint Detection



The controller moves 1 throttle valve on the pump line 2 source radio-frequency generator, wired to the coil 3 bias radio-frequency generator, wired to the chuck 4 backside helium valve

The controller reads 1 chamber pressure gauge 2 optical emission head at the window: density and endpoint 3 self-bias pickup at the chuck 4 wafer temperature probe

What disturbs it 1 polymer film building on the walls after a wet clean 2 drift of the radio-frequency match 3 leaking chuck helium seal

Notice the bare chamber walls: after a wet clean they slow the etch of the first wafers while the plasma loops hold their setpoints, and only the optical head, which watches the glow change as the film clears, knows when to stop. Illustration, not to scale.

IN PLAIN TERMS **Knowing When to Stop Etching**

The main idea. Chips get their tiny patterns by etching. A glowing gas eats away a thin film wherever a protective mask leaves it bare, and four automatic controls keep the gas, the glow and the wafer temperature steady.

The challenge. After the chamber is cleaned, its bare walls slow the etching, and none of the four controls can see that. An etch stopped by a timer then leaves the film only partly removed on the first wafers.

Why it matters. A light sensor that watches the glow change as the film runs out knows when to stop, and it saves wafers that better tuning cannot.

THREE WAYS TO READ THIS CHAPTER

Explore	12 min	The first page, the fact box of Section 7.1, the history box of Section 7.2, the try-it boxes and the fact box of Section 7.6, and the key points
Study	50 min	The whole chapter with its pause points and the twelve exercises
Practice	20 min	The first page, Section 7.6, Section 7.7, Section 7.8 and the open challenge of Section 7.9

Table 7.1 Measured open-loop step response of each variable to its own actuator from the design point, with the four loops out of service, for steps of 5 W of bias, 10 W of source power, 0.3 percentage points of throttle and 0.1 Torr of helium, timed from the controller scan that applies the step. The main etch step lasts 86 s.

Actuator	Variable	τ_{63} (s)	τ_{95} (s)
Bias generator	self-bias V_{dc}	0.063	0.147
Source generator	plasma density n_e	0.104	0.285
Throttle valve	pressure P	0.480	1.115
Backside helium	wafer temperature	15.3	45.0

The control question The four loops of an etch chamber settle in a small fraction of the time a wafer spends in it and barely interact. What then decides whether a cassette survives a chamber clean: the loop tuning, the recipe setpoints, a model predictive controller, or a detector that decides when to stop?

Learning objectives

- Justify the pairing of the four chamber loops from the measured gain matrix and its relative gain array, and name the couplings the array cannot see.
- Explain from the rate, sputter and uniformity laws why ion energy and pressure trade etch rate, mask selectivity, uniformity and sidewall angle.
- Compute the over-etch a wafer needs from the clearing transition and the endpoint threshold.
- Explain why a timed recipe loses wafers after a wet clean while the loops hold their setpoints, and why an emission endpoint recovers them.
- Attribute the measured gain of a predictive controller over four loops to tuning or to its interaction terms.

Prerequisites Sections 1.4.2, 1.4.6 (endpoint detection), 1.6.2 (velocity form and loop action), 1.6.3 and 1.6.4; Chapter 6 for the etcher inside a run-to-run loop.

The app https://apmonitor.com/apps/plasma-etch/

See also Volume 6, Chapter 2 for pairing on a triangular gain matrix and Volume 6, Chapter 8 for a predictive controller above a decentralized structure. Volume 8, Chapter 7 for a phase sequencer above two loops, and Volume 8, Chapter 8 for a batch step ended on an estimate of the product.

7.1 Introduction

An inductively coupled plasma etch chamber has four measurements a control system can close on and four actuators to close them with. Chamber pressure goes to a throttle valve, plasma density to the source generator, self-bias voltage to the bias generator, and wafer temperature to the backside helium pressure. All four are measured directly, all four have dedicated actuators, and none of them carries meaningful dead time.

Table 7.1 shows that the slowest loop on the chamber settles in about a third of the time a wafer spends in it, and the fastest settles in under two tenths of a second. From the wafer's point of view the chamber is at steady state throughout. The dynamic control problem, which occupies most of a process control course, is largely solved by the hardware.

What is not solved is which four setpoints to choose. Each of them moves several of the properties the wafer is graded on, the properties trade against each other, and none of the trades has an interior optimum in more than one variable at a time. The emphasis of this application is therefore the reverse of the usual one. The loops are a prerequisite and the recipe is the controller.

The second reversal concerns what counts as a control element. A cassette of wafers is a batch sequence rather than a regulator. Section 7.6.5 measures a case in which loops holding pressure, density and self-bias to four decimal places scrap the first three wafers after a chamber clean, and Section 7.6.6 shows that the remedy is a detector that decides when to stop.

✦ DID YOU KNOW? A Clean Every Thousand Wafers

In a production trench etch studied in 2011, each chamber was stopped for maintenance after about 1000 wafers: its inner walls were cleaned and its ceramic parts and electrodes replaced. Each time the tool was out of service for up to fourteen hours [1].

If the loops are fast and nearly decoupled, a controller that plans all four actuators at once should add little, and the application carries one so that the claim is tested rather than assumed. Section 7.6.7 sets a four-input, four-output model predictive controller against the shipped loops and against the same loops retuned once, and assigns each measured difference either to a term of the controller's model or to tuning.

The chamber teaches a specific set of process dynamics and control principles. Loop pairing on an interacting plant is decided from a measured steady-state gain matrix and its relative gain array rather than from intuition, and the pairing that results is defensible only because the time-scale separation between the loops and the batch is large. Time-scale separation itself is the second principle, because a regulatory layer that is two orders of magnitude faster than the product converts a dynamic control problem into a static optimization problem. Feedforward recipe execution is distinguished from feedback regulation by a single measured disturbance, an unmeasured drift in an unmeasured state motivates inferential and batch-to-batch correction, and a discrete sequencer with an event-driven transition sits above the continuous loops as a supervisory layer. The physical fundamentals are a material balance on the chamber under a pumping speed set by the throttle, an energy balance on the wafer whose only cooling path is the backside helium conductance, and a surface reaction rate that requires ion flux and neutral coverage together. Those are the same conservation laws and the same loop structures that govern the rest of a process dynamics and control course [2–4], applied to a plant whose product is graded once and cannot be reworked.

The behavior matters industrially because an etch chamber processes a wafer whose accumulated value at that step is large, the pattern is carried in a resist that is consumed by the process itself, and a scrapped wafer is scrapped outright. This chapter describes the simulator in enough detail to rebuild it and measures what each control element is worth on a graded cassette where several of them are worth nothing. Section 7.2 reviews the ion-assisted etch mechanism, endpoint detection and the two disturbances the chamber instruments cannot see. Section 7.3 justifies the loop pairing from the measured gain matrix. Section 7.4 gives the chamber model, and Section 7.5 the four loops, the wafer sequencer, the endpoint detector and the predictive controller. Section 7.6 measures the recipe trades, the stabilization step, open-loop and closed-loop operation, chamber seasoning, over-etch, the graded cassette and the predictive controller against the loops. Section 7.7 describes the real sensors, final elements and controller, Section 7.8 states what the model leaves out, and Sections 7.9 and 7.10 draw the lessons.

7.2 Background

Plasma etching became the dominant pattern-transfer step because it is anisotropic where a wet etch is not. How its ion-assisted mechanism was found, and how the end of an etch came to be detected, is what the History box below tells. The synergy of ions and neutrals is what makes a directional ion flux able to select where a chemical etch proceeds, and it is the term that carries the etch rate in the model below. The discharge physics behind the ion flux, the sheath and the self-bias that accelerates ions into the wafer is standard material [5].

Knowing when to stop is a separate problem and it was solved separately. The optical emission of the discharge changes as the film being etched is consumed and the etchant is no longer being used. Endpoint detection is unusual among industrial measurements because it observes the product rather than the equipment, which is what makes it robust to disturbances the chamber instruments cannot see. Section 1.4.6 sets out the timed recipe against the endpoint on the running example of this volume.

Table 7.2 Measured steady-state gain matrix at the design point, all loops out of service, from a central difference over a 0.5 % perturbation of each actuator. The source column moves all four rows and the throttle column three.

	Source (1/W)	Bias (1/W)	Valve (1/%)	Helium (1/Torr)
n_e (1)	8.94×10^{-4}	0	-2.68×10^{-2}	0
V_{dc} (V)	-5.37×10^{-2}	4.00×10^{-1}	1.61	0
P (mTorr)	3.48×10^{-3}	0	−1.53	0
T_w (°C)	1.82×10^{-2}	1.21×10^{-1}	0	−3.50

HISTORY Learning to Aim and to Stop (1979 and 1980)

The first etching processes used liquids, which etch at the same rate in every direction and so limit how small a feature can be [1]. In 1979 an experiment exposed silicon to xenon difluoride gas and to argon ions, separately and together, and found the etch rate with both greater than the sum of the two acting alone. That ion-assisted chemistry goes with the directional profiles of plasma etching [1, 6]. Knowing when to stop followed at once. In 1979 the impedance of an aluminum etch discharge was shown to change at the end of the etch [1], and in 1980 a study measured the lower limit of detection of an optical emission endpoint for polycrystalline silicon and silicon nitride films [7]. Actinometry, which compares an emission line with that of an inert gas added in known amount, was described the same year [1]. Detectors that read thousands of wavelengths at once came decades later [8].

Two of those disturbances appear in this application. The interior of the chamber accumulates a fluorocarbon film that changes the radical recombination rate at the wall, so the first wafers after a clean are etched at a different rate from the wafers that follow at identical actuator positions [9], and the across-wafer distribution of etch rate in a high-density reactor is set by a competition between ion focusing at the edge and radical depletion at the center [10]. Production tools answer both with layers above the regulatory loops, chamber condition models that estimate what the instruments do not measure [11] and run-to-run control that adjusts the recipe between wafers on measured results [12], the subject of Section 1.4.3.

7.3 Interaction Analysis

The chamber is heavily cross-coupled in the open loop, and the natural pairing has to be justified rather than assumed.

Table 7.2 contains three physical couplings worth naming. The source generator raises the pressure, because dissociation turns each admitted molecule into more particles than it started with. The source generator *lowers* the self-bias, because a denser plasma supplies more ions to carry the same bias current at a lower sheath voltage, so turning the source up softens the ion bombardment. And both generators heat the wafer, which the helium loop then has to remove.

Despite that, the relative gain array for the natural pairing is

$$\Lambda = K \odot \left(K^{-1}\right)^{T} = \begin{bmatrix} 1.073 & 0 & -0.073 & 0 \\ 0 & 1.000 & 0 & 0 \\ -0.073 & 0 & 1.073 & 0 \\ 0 & 0 & 0 & 1.000 \end{bmatrix} \tag{7.1}$$

which is within 0.073 of the identity. The pairing is correct, the only real interaction is a mild one between the pressure and density loops through the source generator's contribution to throughput, and the couplings in the V_{dc} and T_w rows are one-directional and therefore invisible to (7.1). Volume 6, Chapter 2 meets the same blind spot in a reverse osmosis train, whose relative gain reads one while

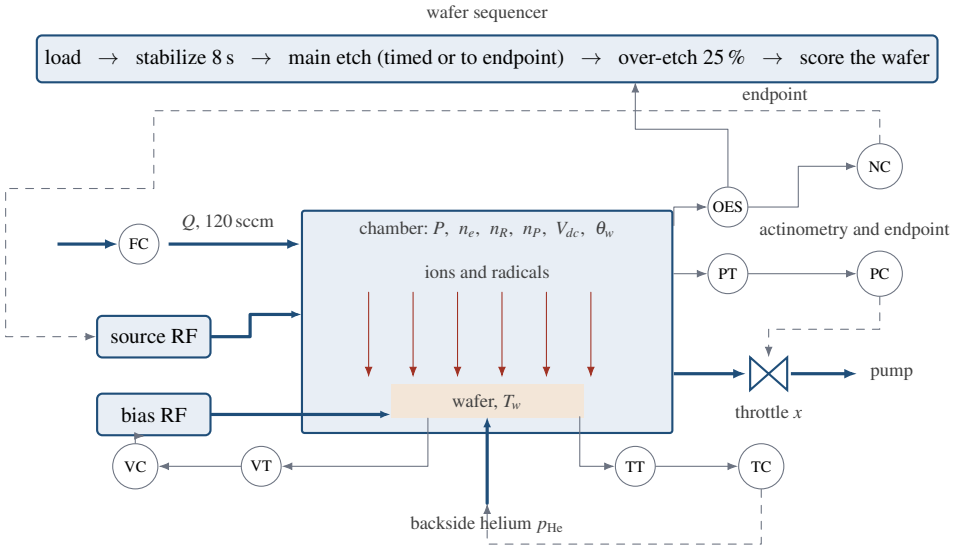


Figure 7.1 The simulated chamber. Four independent single-loop controllers hold pressure on the throttle valve (PC), plasma density on the source generator (NC), self-bias on the bias generator (VC) and wafer temperature on the backside helium (TC). Above them a discrete sequencer runs each wafer through load, stabilize, main etch and over-etch, taking the transition out of the main etch either from a timer or from the optical emission signal.

its loops interact through a triangular gain matrix. Combined with Table 7.1, this establishes that four independent single-loop controllers are the correct structure and that tuning them is a short task rather than a research problem. The remainder of this chapter is about what the setpoints those controllers hold are worth, except for Section 7.6.7, which measures what a controller that does see the one-directional terms gains from them.

7.4 Chamber Model

The chamber, its four loops and the sequencer above them are shown in Figure 7.1. The model carries pressure, plasma density, two neutral species, self-bias, wafer temperature, a chamber wall polymer coverage, three wafer depths and five actuator lags.

7.4.1 Working Units and the Design Point

The clock is in seconds, pressure in mTorr, flow in sccm and etch rate in nm/s. At design the source generator delivers 1200 W, the bias generator 150 W, the throttle valve sits at 28.71 % and the backside helium at 8 Torr. That gives 20.0 mTorr, a normalized plasma density of 1.000, 120 V of self-bias and 60 °C at the wafer, which etch the film at 5.831 nm/s with a mask selectivity of 12.00, an underlayer selectivity of 25.00, 2.200 % non-uniformity and an 88.500° sidewall.

7.4.2 Ion-Assisted Chemistry

The etch rate is the sum of an ion-assisted term and a spontaneous chemical term, inhibited by the passivating species and scaled by the pattern open area,

$$R = \left[k_{IA}\, n_e \sqrt{E_i - E_{\text{th}}}\; \theta_R + k_{CH}\, n_R\, \mathrm{e}^{-E_a/RT_w} \right] \times (1 - k_i \theta_P)\; \phi(A) \quad (7.2)$$

with θ_R and θ_P Langmuir coverages of the etchant and passivant. The ion-assisted term is the mechanism that makes anisotropic etching possible at all, and the synergy it represents was established by Coburn and Winters [6]: ions alone and neutrals alone each etch slowly and together they etch fast.

The mask and the underlayer are attacked by the same ions with different exponents,

$$R_{\text{mask}} = k_m\, n_e\, (E_i - E_m)^{0.85}\, \phi(A), \tag{7.3}$$

$$R_{\text{under}} = k_u\, n_e \sqrt{E_i - E_u}\, (1 - k_p \theta_P)\, \phi(A) \tag{7.4}$$

and the difference between those exponents and the square root in (7.2) is what makes selectivity fall as ion energy rises. The loading factor $\phi(A)$ is common to all three rates, so it cancels in both selectivities. Ion energy is set by the self-bias plus a plasma potential, and the self-bias follows (7.5).

$$V_{dc} = k_V \sqrt{P_{\text{bias}}/n_e} \tag{7.5}$$

A denser plasma supplies more ions to carry the same bias current, so the same bias power develops less voltage, which is where the negative source-to-V_{dc} entry in Table 7.2 comes from [5].

7.4.3 Uniformity and Sidewall Angle

Across-wafer uniformity is the difference of two competing terms,

$$U = U_0 + h_U \left| k_e \left(\frac{P_0}{P} \right)^{0.8} - k_c \left(\frac{P}{P_0} \right)^{0.7} \frac{R}{R_0} \frac{Q_0}{Q} \right| \tag{7.6}$$

with an edge-fast term that dominates at low pressure, where ions focus at the wafer edge, and a center-slow term that dominates at high pressure, where radicals are consumed before they reach the middle. The two cancel at one pressure, which is the interior optimum, and the competition it stands in for is documented in high-density reactors [10]. The reference rate R_0 in (7.6) is the design value of 5.831 nm/s, P_0 and Q_0 are the design pressure and flow, and h_U is a hardware uniformity factor equal to one except in the graded challenge.

Sidewall angle is not optimized at the same place,

$$\alpha = 90 - k_\alpha \left(\frac{P}{P_0} \right)^{0.6} \left(\frac{135}{E_i} \right)^{n_a} \frac{1}{0.5 + \theta_P} \tag{7.7}$$

because pressure scatters the ions and tapers the profile monotonically. Section 7.6.2 measures the resulting conflict.

7.4.4 Wall Polymer Coverage

A single state carries the polymer coverage on the chamber walls,

$$\dot{\theta}_w = \mathbb{1}_{\text{plasma}}\, k_w \left[n_P (1 - \theta_w) - n_R\, \theta_w \right] \tag{7.8}$$

deposited by the passivant and stripped by the etchant, and only while the plasma is lit. It feeds back into the radical balance by shielding the walls, so a chamber that has run wafers all morning has a different radical density from one that has just been cleaned, at identical actuator positions. That is the first-wafer effect [9], and it is the subject of Section 7.6.5.

7.4.5 The Clearing Transition

The film does not clear everywhere at once, because the etch rate is not the same everywhere. The application makes the cleared fraction a logistic in depth whose width is set by the non-uniformity,

$$c(t) = \left[1 + \exp\!\left(-\frac{z(t) - t_{\text{film}}}{\max(0.8,\; 0.60\, U\, t_{\text{film}}/100)} \right) \right]^{-1} \tag{7.9}$$

and the underlayer is only attacked in proportion to c. This one equation carries the whole over-etch argument. A poorly uniform wafer needs more over-etch to clear everywhere, and every second of over-etch is spent cutting into the underlayer wherever it has already cleared.

7.4.6 Initialization and Assumptions

The chamber steady state is algebraic once the pressure is known, and the pressure needs one short iteration because the plasma the source makes changes the throughput. Two inverse calculators return the throttle position for a chosen pressure and the helium pressure for a chosen wafer temperature.

The plasma is a global model with two lumped neutral species and no spatial structure, a normalized electron density rather than an absolute one, and no sheath model beyond the self-bias relation. Integration is fixed-step RK4 at 0.005 s with a 20 ms controller scan, on the engine of Section 1.6.1.

7.5 Controller Design

The control system has three layers, four continuous loops, a discrete wafer sequencer above them, and an endpoint detector that ends one phase of the sequence. A model predictive controller can take the place of the four loops in the regulatory layer, and Section 7.5.4 documents it.

7.5.1 Velocity Form Chamber Loops

Each loop computes an increment to its actuator every control interval, the velocity form of the PID law that Section 1.6.2 sets out [2, 13]. The derivative acts on the measurement rather than on the error, and the four chamber loops carry no derivative term, so the implemented controllers are proportional-integral. Each loop closes on a first-order filtered measurement, with time constants of 0.05 s on pressure and density, 0.02 s on self-bias and 0.30 s on wafer temperature. The error and the three increments are given by (7.10) and (7.11).

$$e_k = a\,(\mathrm{SP}_k - \mathrm{PV}_k) \tag{7.10}$$

$$\begin{aligned} \Delta u_k^P &= K_c\,(e_k - e_{k-1}) \\ \Delta u_k^I &= \frac{K_c}{\tau_I}\,e_k\,\Delta t_c \\ \Delta u_k^D &= -\,K_c\,\tau_D\,a\,\frac{\mathrm{PV}_k - 2\mathrm{PV}_{k-1} + \mathrm{PV}_{k-2}}{\Delta t_c} \end{aligned} \tag{7.11}$$

where $a = +1$ on the density and self-bias loops and $a = -1$ on the pressure and wafer-temperature loops, whose process gains are negative, $\tau_D = 0$, and $\Delta t_c = 20\,\mathrm{ms}$. In the convention of Section 1.6.2 the density and self-bias loops are therefore reverse acting and the pressure and wafer-temperature loops direct acting, which is also what the app's tuning notes call the last two. The increments are combined and the demand is clipped to the actuator range, as shown in (7.12).

$$\begin{aligned} s_k &= s_{k-1} + \Delta u_k^P + \Delta u_k^D \\ c_k &= s_k + I_{k-1} \\ u_k &= \mathrm{clamp}\,(c_k,\ u_{\min},\ u_{\max}) \\ I_k &= I_{k-1} + \Delta u_k^I + \frac{\Delta t_c}{T_t}\,(u_k - c_k) \end{aligned} \tag{7.12}$$

Here s_k, c_k and I_k are the running sum S_k, the output before limiting v_k and the integral I_k of Equations (1.13) to (1.15), with no output bias and a tracking time $T_t = \tau_I$. The actuator ranges are 4 to 100 % of throttle, 0 to 2600 W of source power, 0 to 700 W of bias and 0.5 to 25 Torr of helium. Anti-windup is back-calculation on the integral alone, applied only when the integral gain is nonzero, and the running sum s_k is never truncated. In manual operation and under the open-loop recipe I_k

is set each scan so that $s_k + I_k$ equals the actuator position in force, so the transfer to closed loop is bumpless with no additional logic.

During each wafer exchange the sequencer switches both generators off. With no plasma the measured density and self-bias collapse, so in closed loop the density and self-bias loops hold their last commands and set I_k each scan as in manual operation, and every strike begins from the source and bias power in force before the exchange. The pressure and temperature loops keep regulating, because the throttle and the helium remain in service.

The tuned set used in every closed-loop case has gains of 0.50 %/mTorr on pressure, 600 W per unit of normalized density, 1.20 W/V on self-bias and 0.45 Torr/°C on wafer temperature, with integral times of 0.40 s, 0.20 s, 0.15 s and 15 s. The pressure and temperature integral times are close to the open-loop time constants of Table 7.1, the choice the SIMC rules make for a lag-dominant loop [14], while the density and self-bias integral times are two to two and a half times their τ_{63}. The application also offers an open-loop mode in which the recipe commands generator power and valve position directly, which is how a large fraction of production recipes run, and Section 7.6.4 measures the difference.

7.5.2 The Sequencer

Each wafer runs through four phases. Load takes 6 s with the plasma off, and the new wafer arrives at 23 °C. Stabilize takes 8 s with the plasma lit and the wafer shielded, so the strike transient happens before the product is exposed. Main etch runs either for a fixed time or until the endpoint detector fires, subject to a 150 s backstop. Over-etch runs for a set percentage of the main etch time. The wafer is then scored against seven independent criteria with no partial credit, namely cleared to at least 0.995 of its area, mask remaining above 60 nm, underlayer loss below 12 nm, non-uniformity below 4 %, sidewall not tapered below 87°, sidewall not undercut above 90.2°, and resist never above 95 °C. Volume 8, Chapter 7 runs the same kind of phase sequencer above two loops inside every shot of a molding machine.

7.5.3 The Endpoint Detector

The endpoint detector watches an actinometric emission signal that falls as the film clears and the etch product stops being generated [7]. The modeled signal is $(1 - c) + \varepsilon$ with c the cleared fraction of (7.9) and $\varepsilon = 0.12$ the residual emission after the film is gone, passed through a first-order filter of 0.30 s and compared against a fixed threshold of 0.55. The threshold corresponds to 57 % of the wafer cleared, and the filter lag moves the firing on a design-point wafer to 64 % cleared. Thresholding a filtered emission trace is the classical implementation and remains the production standard even as spectral methods improve [8]. It is not a feedback loop. It has no setpoint, no actuator and no error signal, and Section 7.6.6 measures it as the most valuable control element on the tool.

🕮 **Pause.** Try Exercise 7.3 before reading on.

7.5.4 Model Predictive Control of the Four Actuators

The fourth chamber-control structure replaces the four loops by one receding-horizon plan for the throttle, the two generators and the backside helium, the multivariable form that dominates industrial practice [15, 16]. Section 1.6.3 gives the general form of the option, and this section gives what is particular to the chamber. Real-time feedback control of a reactive ion etcher on its measured plasma variables is documented in [17]. The sequencer, the endpoint detector, the recipe and every actuator limit are unchanged. While the plan acts, the four PI loops set I_k each scan so that their output equals the applied command, as in manual operation, which makes a hand-over in either direction bumpless.

The controller's model is identified rather than derived. With the plasma held, the sensor noise off and the actuators at the design point, each actuator is stepped by 5 % of its design value in each direction (1.4 percentage points of throttle, 60 W of source, 7.5 W of bias and 0.4 Torr of helium), the two responses are averaged, and every input-output pair is fitted with a second-order lag and a

Table 7.3 Identified elements of the predictive controller's model, from ±5 % steps of each actuator about the design point with the plasma held. Residual is the root-mean-square fit error as a share of the response.

Input	Output	K	τ_1 (s)	τ_2 (s)	θ (s)	Resid.
throttle	P	−1.542 mTorr/%	0.311	0.167	0.0435	0.27 %
throttle	n_e	−0.0269 1/%	0.291	0.197	0.0811	0.30 %
throttle	V_{dc}	1.61 V/%	0.299	0.180	0.0802	0.29 %
source	P	3.483 mTorr/kW	0.309	0.068	0.0294	0.30 %
source	n_e	0.893 1/kW	0.087	0.049	0.0237	0.71 %
source	V_{dc}	−53.6 V/kW	0.075	0.042	0.0319	0.82 %
source	T_w	18.19 °C/kW	15.00	–	0.0398	0.00 %
bias	V_{dc}	0.400 V/W	0.038	0.038	0.0102	0.48 %
bias	T_w	0.121 °C/W	15.00	–	0.0398	0.00 %
helium	T_w	−3.507 °C/Torr	15.04	0.391	–	0.00 %

Exactly zero: bias to P and n_e, helium to P, n_e and V_{dc}, throttle to T_w.

dead time,

$$G_{ij}(s) = \frac{K_{ij}\,\mathrm{e}^{-\theta_{ij}s}}{(\tau_{1,ij}s+1)(\tau_{2,ij}s+1)} \tag{7.13}$$

where K_{ij} is the steady gain of output i to input j, $\tau_{1,ij}$ and $\tau_{2,ij}$ are the two time constants and θ_{ij} is the dead time. Ten pairs respond and six measure exactly zero, as listed in Table 7.3. The pressure, density and self-bias are fitted on the instrument readings and the wafer on its true temperature, which omits the 0.30 s probe filter against a 15 s lag. The gains equal those of Table 7.2 to its three figures, each residual is at most 0.82 % of the response against 0.56 to 3.23 % for a first-order fit, and an independent set of step tests read at every 20 ms scan repeats the ten gains within 0.26 %.

Every dead time is shorter than one 0.1 s prediction step, so each element is realized as a chain of first-order lags with the dead time written as one more lag of its own length. The three wafer elements fit time constants of 15.00, 15.00 and 15.04 s, the single heat balance $C_w/h_{\mathrm{He}} = 60/4 = 15\,\mathrm{s}$ of the chuck, so they share one thermal state x_T with a heat-load disturbance w,

$$\tau_T\,\dot{x}_T = -x_T + w + \sum_j K_{Tj}\,v_j \tag{7.14}$$

where $\tau_T = 15\,\mathrm{s}$, the sum runs over the source, bias and helium, K_{Tj} are the three wafer gains of Table 7.3 and v_j is input j after its element's short lags. The result is a linear model of 26 states in deviation variables about the design point, discretized exactly under a zero-order hold at 0.1 s for the plan and at 20 ms for the estimator. Its numbers are frozen and it reads no chamber parameter or disturbance, so a change made to the simulated chamber reaches the controller only through the instruments, as model mismatch. Three models are therefore kept apart, the fifteen-state nonlinear plant of Section 7.4, the four filtered instruments, and this 26-state identification.

Offset-free tracking comes from a disturbance model in the estimator [18]. At every 20 ms scan, under every structure, the commands applied over the last scan drive the model, with zero source and bias power while the RF is off, and the four instruments correct it through steady-state Kalman gains. The pressure, density and self-bias each carry a constant output disturbance modeled as a random walk (0.045 mTorr, 0.0022 and 0.22 V per scan, against measurement noise of 0.2 mTorr, 0.01 and 1 V), which gives a gain of 0.200 per scan and an error time constant near 0.09 s. On the wafer the thermal state and the heat load of (7.14) are corrected together with gains of 0.618 and 0.309, so a reading is taken up within one scan while a heat load, which appears only through the 15 s lag, is estimated over about 30 s. While the RF is off the density and self-bias corrections are held, because there is no plasma to measure.

Every five scans (0.1 s, counted in whole scans) and at once whenever the RF strikes or goes off, the controller solves

$$\begin{aligned} \min \; & \sum_{k=1}^{N} \sum_{i=1}^{4} \left(\frac{\hat{y}_{i,k} - r_i}{b_i} \right)^2 + \sum_{j=0}^{N_b-1} \sum_{m=1}^{4} \left(\frac{\Delta u_{m,j}}{0.1\,\Delta_m} \right)^2 + \rho\, \varepsilon^2 \\ \text{s.t. } & u_{m,\min} \le u_{m,j} \le u_{m,\max}, \quad \left| \Delta u_{\mathrm{src},j} \right| \le 300\,\mathrm{W}, \\ & \hat{y}_{T,k} \le T_{\max} - 3\,^{\circ}\mathrm{C} + \varepsilon, \quad \varepsilon \ge 0 \end{aligned} \tag{7.15}$$

over $N = 50$ steps of 0.1 s, a 5 s horizon, with the inputs held constant across $N_b = 8$ blocks whose lengths in steps are $(1, 1, 2, 2, 4, 6, 10, 24)$. Here $\hat{y}_{i,k}$ is the predicted measurement including its disturbance estimate, r_i the setpoint, b_i the loop's tolerance band (0.6 mTorr, 0.04, 4 V and 3 °C), $\Delta u_{m,j}$ the move of actuator m at block j, measured for $j = 0$ from the command in force, Δ_m the actuator span (96 %, 2600 W, 700 W and 24.5 Torr), $T_{\max} = 95\,^{\circ}\mathrm{C}$ the resist limit, ε one slack over the whole horizon and $\rho = 10^4\ 1/\mathrm{K}^2$. A band's worth of error at one step therefore costs as much as a move of a tenth of the actuator's span. These weights were fixed by that rule before any run, and one pass at 0.3 and 3 times the move weight, made before the rate limit was added, changed the wafer result of the new scenario by less than 1 %. The actuator box is hard and equal to the clamps of the loops. The source rate limit is hard and applies at every block boundary, because a real RF generator is ramped at ignition and a match network tuning into a plasma that has just lit can reflect power and arc. The plant has no reflected-power or arc model, so the benefit of that row is visible only as the move it removes, which Section 7.6.7 measures. The wafer ceiling at 92 °C is soft. While the RF is off, the blocks before the strike hold the source and bias commands and the density and self-bias terms carry no weight. The decisions are scaled by the actuator spans before the shared dual active-set solver receives the condensed quadratic program, because watts against percent of throttle would otherwise spread its Hessian over eight decades.

The only future event the plan is given is the strike that ends each wafer exchange, which the sequencer schedules from its own load time. With this preview on, the plan knows that the source and bias power reach the chamber at that instant, and with it off the plan holds the present RF state over the horizon and learns of the strike when it happens. Recipe setpoint changes are previewed by neither structure. For the comparison the loops receive the same signal as a feedforward that adds the identified dissociation term to the throttle, $K_{P,\mathrm{src}} P_{\mathrm{src}} / |K_{P,x}|$ or 2.7 percentage points at 1200 W, from 0.3 s before each strike. That feedforward is a row of the comparison script, not an option of the application.

Before a plan reaches the actuators it is checked. The solver's status must be optimal, every number it returns (first move, later moves, predicted outputs, cost and slack) must be finite, each planned command must sit within its box, the source moves must respect the rate limit, and the generators must stay held while the RF is off. Any other outcome is rejected with its reason and marked in the run record, and the four loops, which have tracked every applied command, continue from the same commands. The next accepted plan takes the actuators back over a 0.3 s ramp. None of the roughly 9010 plans of any 900 s cassette reported here failed these checks. Each plan costs about 1.0 ms in Node.js, and in the browser, on the page's own frame loop, the solves take 1.6 to 3.0 % of one core at real time and 5.8 to 7.3 % at five times real time, over two measurements with different amounts of other work on the computer.

7.6 Results

7.6.1 Ion Energy Against Mask Selectivity

Table 7.4 is the central trade and it is not a two-way one. Raising the ion energy from 95 to 185 eV makes the etch 42 % faster, the uniformity 38 % better and the sidewall 0.93° closer to vertical, and it costs 35 % of the mask selectivity because (7.3) sputters the mask with a higher exponent than (7.2) etches the film.

Table 7.4 Closed-loop steady state against self-bias setpoint, all other setpoints at design. Every column but the mask selectivity improves with bias.

V_{dc} (V)	E_i (eV)	R (nm/s)	Mask sel.	U (%)	α (°)
80	95.0	4.828	15.34	2.716	87.942
120	135.0	5.831	12.00	2.200	88.500
170	185.0	6.852	9.99	1.675	88.870

Table 7.5 Closed-loop steady state against pressure setpoint, all other setpoints at design and the wall coverage at its design value of 0.5. Uniformity reaches its floor of 1.20 % at 22.8 mTorr and the sidewall has no interior optimum.

P (mTorr)	R (nm/s)	U (%)	α (°)	Mask sel.
12	3.613	5.919	88.601	7.44
16	4.867	3.840	88.543	10.02
20	5.831	2.200	88.500	12.00
23	6.404	**1.256**	88.468	13.18
26	6.881	2.211	88.435	14.16
34	7.864	4.450	**88.337**	16.18

Scenario 8 runs that trade through the wafer sequence. At 220 V the main etch finishes in 63.8 s instead of 85.9 and leaves 230.2 nm of mask instead of 247.4. At 75 V it takes 107.7 s, leaves 260.6 nm of mask, and pushes the sidewall to 87.84°, which is within a degree of the specification floor. All three pass. The choice among them is throughput against mask budget against profile margin, and it is a business decision rather than a control decision.

7.6.2 Pressure, Uniformity and Sidewall Angle

The uniformity column of Table 7.5 is a valley whose floor lies 2.8 mTorr above the design point, where the edge and center terms of (7.6) cancel, and the sidewall column is a monotone slope in the opposite direction, so no pressure optimizes both. The rate and the mask selectivity both improve monotonically with pressure, so the incentive to run at high pressure is direct and the reason not to is indirect.

Scenario 9 measures the cost of moving in the other direction. Moving from 20 to 13 mTorr for a better sidewall takes the non-uniformity to 5.34 %, above the 4 % specification, and scraps both wafers etched there. The main etch also lengthens from 85.9 to 127.3 s, so the decision costs throughput as well as yield.

> **TRY IT IN THE APP** **9 · Uniformity Against Sidewall**
>
> Press Run and let the cassette step its pressure setpoint. At 13 mTorr the sidewall is straighter, but the non-uniformity reaches 5.34 %, above the 4 % grade, and both wafers etched there are scrapped. Read the main etch time as well: it stretches from 85.9 to 127.3 s.

7.6.3 The Stabilization Step

Scenario 5 deletes the eight-second stabilization step, which is worth seven percent more wafers an hour (one every 113.6 s instead of 121.6 s). The transient is not removed by that change. It is relocated.

With the stabilization step in place, the pressure holds 20.00 mTorr and the self-bias 120.0 V through the whole main etch, and the etch rate varies only between 5.545 and 5.958 nm/s as the wafer warms. With it deleted, the strike lands in the main etch. The self-bias overshoots to 203.3 V while the density is still rising, then dips to 65.5 V, the pressure rises to 22.38 mTorr, and the rate climbs

Table 7.6 Eight wafers after a wet clean, with the wall coverage and the etch rate read at the end of each main etch. The loop setpoints never move.

Wafer	θ_w	R (nm/s)	Cleared	Result
1	0.078	5.393	0.9288	scrapped
2	0.152	5.464	0.9742	scrapped
3	0.213	5.526	0.9897	scrapped
4	0.264	5.577	0.9954	good
8	0.393	5.714	0.9995	good

from near zero. By the time the pressure is within 0.6 mTorr of setpoint, 1.3 s later, the top 5.6 nm of every feature has been etched at the wrong rate and ion energy, where the sidewall shows it.

The application still passes those wafers. Its scoring computes uniformity and sidewall from the final chamber state rather than integrating the profile over depth, so a defect confined to the top of the feature is not represented. The traces show the excursion clearly and the scorecard does not see it, which is a limitation of the model and is recorded as such in Section 7.8.

7.6.4 Open-Loop and Closed-Loop Chamber Control

Scenario 6 detunes the RF match at 130 s so that ten percent of the forward power reflects. The generator continues to report 1200 W.

Under the open-loop recipe the plasma density falls from 1.000 to 0.900, the etch rate at the end of each main etch from 5.834 to 5.209 nm/s and, less obviously, the self-bias *rises* from 120.0 to 126.5 V, because (7.5) has fewer ions carrying the same bias current. The recipe has changed both the rate and the ion energy while every reported actuator value is unchanged.

Putting the chamber loops in service at 400 s raises the source command to 1333 W, returns the density to 1.0000 and the self-bias to 120.0 V within 5 s. The pressure loop opens the throttle from 28.70 to 29.01 % to absorb the extra dissociation, which leaves the rate 0.6 % below that of the undisturbed closed-loop chamber. That is the argument for closed-loop chamber control in one measurement. Match drift, mass flow calibration, wall condition and gas purity all pass straight through an open-loop recipe and are absorbed by a closed-loop one, which is why production tools carry chamber-condition models beside their loops [11].

Pause. Try Exercise 7.2 before reading on.

7.6.5 Chamber Seasoning After a Wet Clean

Scenario 7 starts with a freshly cleaned chamber and runs a cassette with a timed recipe, an 86 s main etch and 12 % over-etch.

> **TRY IT IN THE APP** **7 · Chamber Clean, and What the Timer Does Not Know**
>
> Load the scenario and press Run. Watch the wall-polymer trace on the chamber strip climb while the pressure, density and self-bias tiles hold their setpoints. Then read each wafer's verdict note as it is graded: wafers 1, 2 and 3 are scrapped because the film has not cleared, and wafer 4 is good, with nothing on the tool changed.

Table 7.6 loses three wafers to a variable that is not measured, not controlled and not displayed anywhere on the tool. The wall polymer coverage of (7.8) climbs from 0.078 to 0.393 over eight wafers, the etch rate rises 6 % with it toward the design value of 5.831 nm/s that the model reaches at a coverage of 0.5, and the recipe's fixed 86 s is too short until the chamber has seasoned. The pressure, density and self-bias loops read 20.000 mTorr, 1.0000 and 120.00 V at the end of every main etch, and the temperature loop repeats the same warm-up from 42.6 to 66.4 °C on every wafer, which is the

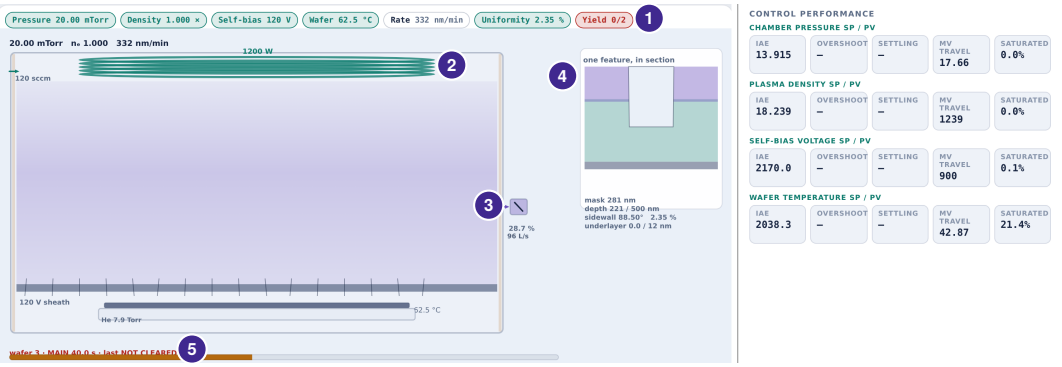


Figure 7.2 Scenario 7 at 274.8 s (4.58 min), 40.0 s into the main etch of the third wafer out of a clean chamber, cropped to the chamber view and, right of the rule, the loop rows of the Learn panel. The four loops read 20.00 mTorr, 1.0000 normalized density, 120 V of self-bias and 62.5 °C at the wafer (tiles along the top). (1) The yield tile. (2) The coil, at 1200 W of source power. (3) The throttle valve on the pump line. (4) The feature cross-section. (5) The wafer sequencer's status line. The integrated absolute errors are 13.9 mTorr s on pressure and 18.2 s on density, while the self-bias and temperature integrals (2170 V s and 2038 °C s) accumulate during each exchange, when there is no plasma to bias, and while each wafer warms from 23 °C. The yield is zero of two (1) and the last wafer is reported as not cleared (5). The feature cross-section (4) shows the etch 221 nm into a 500 nm film with 281 nm of the 300 nm mask left.

point: the loops are controlling the plasma and the plasma is not the whole plant. With the main etch ended by the emission detector instead, the same eight wafers all pass, and the main etch it chooses shortens from 93.6 s on the first wafer to 87.8 s on the eighth as the walls season.

🕮 **Pause.** Try Exercise 7.1 before reading on.

> ✦ **DID YOU KNOW? Close Rates, Opposite Verdicts**
>
> From the simulator: in Table 7.6 the first wafer after the wet clean etches at 5.393 nm/s, clears 0.9288 of its area and is scrapped. The fourth etches at 5.577 nm/s, clears 0.9954, just past the 0.995 grade, and is good. The pressure, density and self-bias read the same on both.

7.6.6 Over-Etch and the Graded Cassette

Over-etch has a two-sided optimum that scenario 10 walks across. At 5 % the wafers do not clear, at 0.987 of the wafer against a 0.995 threshold. At 30 % they clear and lose 6.2 nm into the underlayer. At 75 % they clear and lose 15.2 nm, above the 12 nm limit. Three good wafers out of six, failing at both ends for opposite reasons.

The graded challenge combines that with everything else: a cassette after a wet clean (wall coverage 0.05), hardware with $h_U = 1.15$ in (7.6), a chuck seal already leaking at 80 % of nominal, a match that drifts to 93 % efficiency at 200 s, a product with 1.45 times the open area entering the cassette at 450 s and the seal degrading to 62 % at 700 s. It loads with an open-loop recipe, a timed 86 s main etch, 10 % over-etch and loop gains about an eighth of the tuned set, and it scraps all eight wafers. The challenge is met when, after 900 s, no more than one wafer has been scrapped and at least five are good.

Table 7.7 is the chapter's main result and it is ordered by increasing value. Retuning the four loops is worth nothing, because in the loaded recipe they are not in service and the tuning constants are inert. Closing the four chamber loops is also worth nothing on its own, with the loaded gains or the tuned set, because a timed main etch in a fresh chamber is too short whatever the plasma does.

Table 7.7 The graded cassette. Each row adds one element to the recipe as loaded, not to the recommended one.

Change from the loaded recipe	Wafers good	Worst cleared
none	0 of 8	0.001
retune all four loops	0 of 8	0.001
close the four chamber loops	0 of 8	0.006
close the loops with the tuned set	0 of 8	0.029
raise over-etch to 32 %	5 of 7	0.965
switch to the OES endpoint	**7 of 7**	**0.996**
close the loops and use the endpoint	**7 of 7**	**0.998**

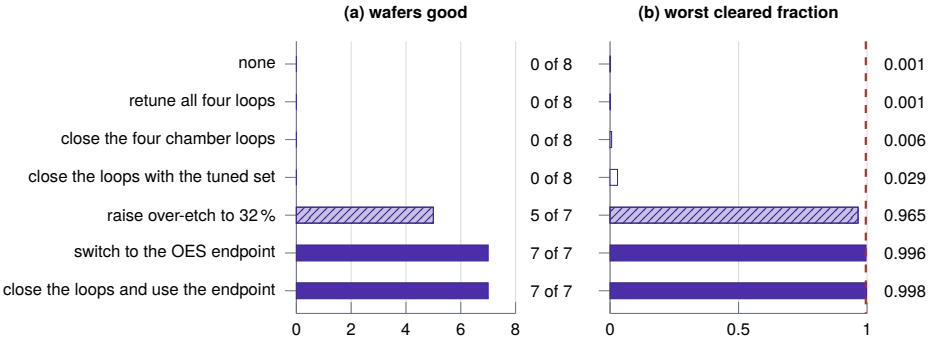


Figure 7.3 The graded cassette of Table 7.7: (a) good wafers of those graded in 900 s, and (b) the worst cleared fraction, against the 0.995 grade (dashed). Open bars change only the loops, the hatched bar the over-etch and the solid bars the end of the main etch. Notice that no change to the loops moves either panel, while the emission endpoint alone takes the worst wafer past the grade.

Enlarging the over-etch is worth five wafers and cannot fix the two that need the most, because a fixed percentage of a fixed main etch is still a fixed time. Figure 7.3 draws the seven rows as bars.

The single most valuable change on the tool is the endpoint detector, worth seven of seven on its own and enough by itself to meet the challenge, and the reason is visible in the times it produces. Under the recommended recipe (closed loops, the emission endpoint, 32 % over-etch and setpoints of 21 mTorr, 108 V and 56 °C), all six wafers graded in 900 s are good and the main etch runs for 95.0, 93.6, 92.5, 101.2, 103.5 and 102.2 s on successive wafers, a spread of 12 %, as the chamber seasons and the product open area changes. The timer in the loaded recipe holds 86 s in every case. The recommended settings give the same run whether they are chosen before Run or carried through a Reset.

Closing the loops as well gives the same seven good wafers at a worst clearance of 0.998 against 0.996. Neither element substitutes for the other's failure mode. The loops reject disturbances inside the chamber and the detector rejects everything the chamber cannot see, including the wafer itself.

7.6.7 Predictive Control Against the Four Loops

Scenario 11 is the cassette of scenario 4 run by the predictive controller of Section 7.5.4, with the strike preview on, a timed main etch and a recipe step of the self-bias setpoint from 120 to 150 V at 300 s. Bias power heats the wafer, which is a coupling that (7.1) cannot see because it runs in one direction only. The comparison of Table 7.8 runs scenario 11, scenario 8 (self-bias steps to 220 and then 75 V, emission endpoint) and scenario 5 (the strike inside the main etch) for 900 s each, at the same seed with the sensor noise off, the same instruments and the same sequencer, and grades

Table 7.8 Predictive control against the four loops. Seconds per wafer outside the loop bands from the true outputs, seven good wafers per 900 s run in every row. Scenario 11 steps the self-bias setpoint, scenario 8 (Sc. 8) runs the bias trade and scenario 5 strikes the plasma inside the etch.

Structure	Scenario 11				Sc. 8	Scenario 5		Source
	T_w out (s)	all out (s)	all out, lit (s)	T_w peak (°C)	T_w out (s)	peak $\lvert\Delta P\rvert$ (mTorr)	all out (s)	travel (W)
Loops, shipped tuning	37.08	37.18	48.31	67.5	42.86	2.38	43.98	482
Predictive controller	3.54	3.56	16.54	62.0	3.74	0.78	19.77	1071
Loops retuned once (SIMC)	4.14	4.16	13.50	61.9	4.89	2.26	15.11	1309
Predictive, diagonal model	4.18	4.20	13.74	62.8	12.13	2.13	15.28	1132
Predictive, strike preview off	3.54	3.56	16.44	62.0	3.74	0.55	19.70	1088
Predictive, no constraint rows	3.85	3.88	12.51	62.0	4.00	0.33	14.06	3553
Predictive, no source rate limit	3.49	3.52	16.47	62.0	3.75	0.25	19.69	3749
Loops with strike feedforward	37.05	37.16	47.53	67.5	42.84	1.46	43.18	432

The retuned loops have gains of 1.01 %/mTorr, 1294 W per unit density, 2.44 W/V and 11.1 Torr/°C with integral times of 0.39, 0.11, 0.057 and 1.56 s. Source travel is the total source command movement per wafer in scenario 11. The predictive rows solve about 9010 plans per run at a mean of 1.0 ms each in Node.js (0.43 ms with the diagonal model, measured with other work on the computer), and none is rejected. "No constraint rows" removes the actuator box, the source rate limit and the wafer ceiling and clips the first move afterward. Without the rate limit the plan of scenario 11 is the one shipped before that limit was added.

seven wafers per run. Every wafer of every row is good. The figure of merit is the time per wafer that each true output spends outside its loop's tolerance band while the wafer is exposed (main etch and over-etch), summed over the four outputs, and the same while the plasma is lit with a wafer in the chamber, which adds the stabilization step. The rows are the shipped loops, the predictive controller as scenario 11 ships it, the same loops retuned once by the SIMC rules [14] from the four diagonal elements of Table 7.3, and five counterfactuals, each with one mechanism of the controller removed or one of its signals given to the loops. No case is retuned for any row.

The predictive controller cuts the time the wafer spends outside its 3 °C band from 37.08 to 3.54 s per wafer on scenario 11 and from 42.86 to 3.74 s on scenario 8, and its hottest wafer is 62.0 °C against 67.5 °C, as listed in Table 7.8. Almost all of that gain is tuning. The shipped wafer loop is slow (0.45 Torr/°C with a 15 s integral time), and the same four loops retuned once by a standard rule reach 4.14 and 4.89 s, which is 98 and 97 % of the controller's improvement. What the bias-to-wafer interaction term adds is the remainder, 0.60 and 1.15 s per wafer, and it is obtained with a seventh and a quarter of the retuned loop's helium travel (24 against 170 and 39 against 155 Torr per wafer), because the plan raises the helium when the bias power rises instead of waiting for the wafer to warm. The diagonal-model row overstates the term on scenario 8 (12.13 s), and most of that loss is its own move weight, since the same diagonal plan reaches 8.41 and 6.01 s at 0.3 and 0.1 times the weight while the full plan stays at 3.74 s.

The strike pressure has a different source. With the stabilization step deleted, the shipped loops let the pressure move 2.38 mTorr at each strike and the predictive controller 0.78 mTorr. That gain belongs to the source-to-pressure term, because the same plan on a diagonal model leaves 2.13 mTorr, and the loops given the preview's signal as a throttle feedforward reach 1.46 mTorr. The re-plan at the moment of the strike tells the plan the generators are on, so the term acts with or without the preview. The preview itself now costs pressure, 0.78 against 0.55 mTorr with it off. Knowing the source will ramp, the plan closes the throttle from 25.8 to 22.9 % in the last half second before the strike, raising the pressure about 0.85 mTorr so that the higher pressure brings the density up sooner, and the strike then lands on a chamber already above setpoint. Before the source rate limit was added the preview helped instead (0.25 against 0.57 mTorr). On scenario 11 the preview changes no measured second, and the scenario keeps it on as it was fixed before these results.

The rate limit exists because of what the unconstrained plan did at ignition. Its linear model starts the density from zero power and knows nothing of ignition, so at every strike it commanded 2432 W

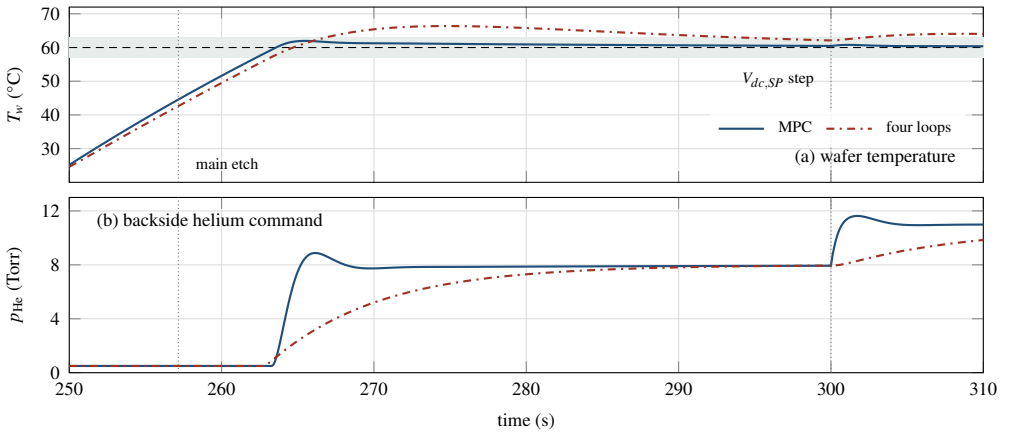


Figure 7.4 Scenario 11 in the 60 s before $t = 310$ s, the third wafer warming from 25 °C into its main etch and the self-bias setpoint stepping from 120 to 150 V at 300 s. (a) Wafer temperature, with its setpoint (thin dashed) and its band (shaded), and (b) backside helium command, under the predictive controller (solid) and the four loops (dash-dotted). Dotted verticals mark the start of the main etch and the setpoint step. At 310 s the wafer reads 60.37 °C with 10.99 Torr of helium under the plan, and 64.05 °C, outside its band, with 9.84 Torr under the loops.

of source for one decision and then 924, 1101 and 1328 W, which sent the true density to 1.42 (the actinometry read 1.19) and reached the 2600 W clamp at every strike once the match had drifted to 0.85. With the limit of 300 W per decision the first move is 1500 W, the density peaks at 1.15, and the source travels 2.2 times as far as under the loops instead of 7.8 times (1071 and 3773 W per wafer in scenario 5 against 482). The limit costs 0.05 s of wafer time on scenario 11 and the strike pressure discussed above.

The plan also decouples. On scenario 3 with the plasma held, a density step from 1.00 to 1.10 drags the self-bias 1.35 V under the predictive controller, 2.16 V under the loops and 5.98 V under the diagonal plan, and a 2 mTorr pressure step is inside a quarter of its band in 0.33 s against 0.90 s under the loops. The diagonal plan matches that speed (0.32 s), so speed comes from the model and the 0.1 s plan, while the smaller cross-excursions come from the interaction terms.

The instant that decides the wafer comparison is shown in Figure 7.4. Both structures warm the third wafer from 25 °C with the helium at its 0.5 Torr floor and reach the band within a second of each other (262.3 and 263.2 s), after which the loops overshoot to 66.4 °C at 274.7 s while the plan stays below 62.0 °C. When the self-bias setpoint steps, the plan raises the helium at once and the wafer peaks at 60.78 °C, while the slow loop waits for the temperature error and the wafer reaches 64.07 °C at 308.8 s.

The losses are as measured. Over the lit time, which includes the warm-up in the stabilization step, the controller spends 16.54 s per wafer outside the bands against 13.50 for the retuned loops and 12.51 for itself without its constraint rows. Knowing the helium is at its floor while a cold wafer warms, the plan heats the wafer with the only other actuators that can, the source and bias power, and holds the density and self-bias up to a band high for several seconds. With the stabilization step deleted that trade falls inside the etch, where the density spends 4.69 s per wafer outside its band against 0.77 under the loops and 0.31 without the rows, and scenario 5's total of 19.77 s loses to the retuned loops' 15.11 s. Sensor noise passes through at about twice the loops' amplitude, a peak pressure error during etching of 0.446 against 0.224 mTorr with the noise slider at one and 1.336 against 0.678 mTorr at three on scenario 4. On scenario 7, with a timed recipe after a chamber clean, the controller grades four good wafers of eight against five, because it holds the wafer at its 60 °C setpoint where the loops overshoot to 66 °C, so the chemical rate is lower and a timer calibrated

Table 7.9 Worst case over every scenario run for 900 s under each structure. Hottest wafer and good wafers of those graded.

Scenario	Four loops	Predictive
4, cassette	66.4 °C, 7 of 7	62.0 °C, 7 of 7
5, stabilization deleted	66.4 °C, 7 of 7	62.0 °C, 7 of 7
6, open-loop recipe[a]	66.9 °C, 7 of 7	66.6 °C, 7 of 7
7, chamber clean, timed	66.4 °C, 5 of 8	62.0 °C, 4 of 8
8, rate against selectivity	76.1 °C, 7 of 7	62.6 °C, 7 of 7
9, uniformity against sidewall	67.4 °C, 4 of 6	62.0 °C, 4 of 6
10, over-etch	66.4 °C, 3 of 6	62.0 °C, 3 of 6
11, predictive control	67.5 °C, 7 of 7	62.0 °C, 7 of 7
12, challenge as loaded	80.2 °C, 0 of 8	62.6 °C, 0 of 8
12, suggested setpoints	66.4 °C, 6 of 6	58.9 °C, 6 of 6
4, chuck seal at 0.40	78.0 °C, 7 of 7	65.4 °C, 7 of 7

[a]The structure named runs from $t = 0$, and the scenario's own event puts the four loops in service at 400 s.

under the loops falls short. The helium also sits at a limit longer, 112.5 against 108.3 s on scenario 4 and 773 against 435 s with the chuck seal at 0.40 of nominal.

No run of Table 7.9 brings any wafer near the 95 °C resist limit, the predictive controller's hottest wafer is the cooler of the two in every scenario that runs wafers and at every disturbance slider of scenario 4 set to its limit, and no plan is rejected anywhere. The match at 0.70, the flow calibration at 0.85 and 1.15, the wall loss at 1.6 and the hardware uniformity factor at 2.2 leave both structures with seven good wafers of seven. The two failures that remain at their limits, a doubled open area and a reduced etchant fraction, leave every wafer uncleared under both structures. The predictive controller is less favorable on scenario 7, by one wafer, and under sensor noise, as described above.

The cost leaves the browser at 60 frames per second throughout. Each plan takes 1.1 to 3.0 ms on average in Chrome, ten plans per simulated second take at most 7.3 % of a core at the top speed, and no frame reached the 25 ms stepping guard that would slow the run (the longest frame stepped for 16.8 ms). The four loops step at five times real time in under 1.0 % of a core.

7.6.8 The Running Application

The application as it runs is reproduced in Figure 7.2. It runs in a browser with no installation and computes everything on the client, and offers manual operation, an open-loop recipe, the four closed loops and the predictive controller. Each run exports as a comma-separated file carrying the four loops, the two neutral densities, the ion energy, the wall polymer coverage, the etch rate, both selectivities, the uniformity, the sidewall angle, the etched depth and the phase, so the recipe trades of Tables 7.4 and 7.5 can be reproduced offline. Manual operation offers two helpers that work on the tool's declared model of a seasoned chamber rather than on the chamber's true state. The first sets the four actuators for the recipe setpoints, through the two inverse calculations of Section 7.4, and so returns the chamber to the design point from which the steps of Table 7.2 can be repeated by hand. The second sets the main etch time from the rate that model predicts.

7.7 Technology in Practice

The loops of Section 7.5 read a pressure gauge, an optical emission spectrometer, a voltage sensor on the bias line and a wafer temperature probe, and they move a throttle valve, two radio-frequency generators and a helium valve. Figure 7.5 draws them with their tags.

What Measures It

Both the density loop and the endpoint detector read the light of the plasma, the first through actinometry and the second through the emission signal of Section 7.5.3.

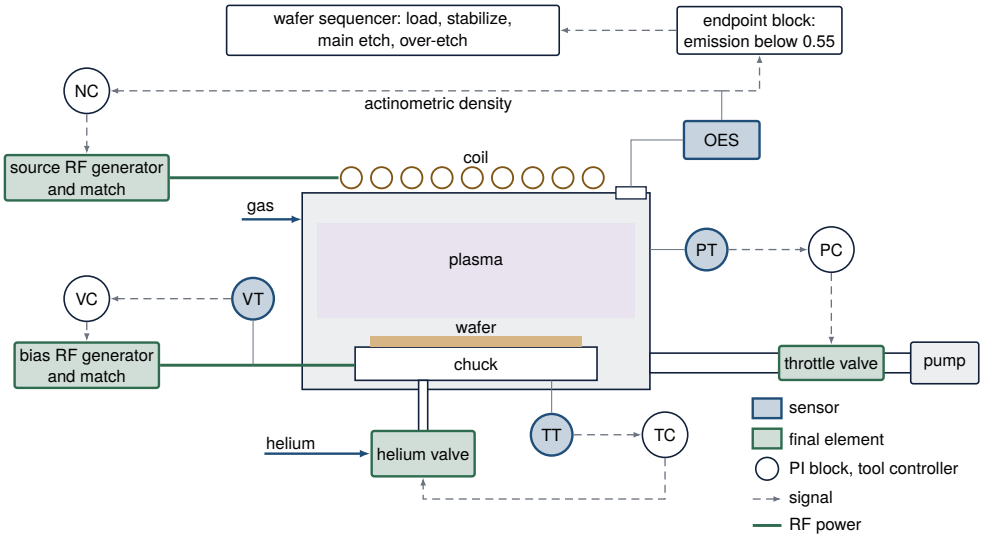


Figure 7.5 The four chamber loops of Figure 7.1 drawn with their hardware. The optical emission spectrometer (OES) at the viewing window serves twice: its actinometric density feeds the density loop NC, and its filtered emission ends the main etch in the wafer sequencer. All four PI loops, the sequencer and the endpoint block run in the etch tool's controller, scanned every 20 ms. Notice that the endpoint signal moves no actuator: it ends a step of the recipe.

TECHNOLOGY Sensor: An Optical Emission Spectrometer

Light from the plasma passes a window and a lens onto a detector, from a filtered photodiode that reads one wavelength to a spectrometer that records the whole spectrum at regular intervals. Each species emits at its own wavelengths, so the light follows the gas composition in real time without touching the plasma. The intensity also depends on the electrons that excite the gas, which is why actinometry compares a line with that of a reference gas. The integration time filters the signal: a longer one gives less noise and a slower reading. Deposits cloud the window and weaken the signal wafer after wafer, and a spectrometer that is knocked can lose its calibration [1].

The 0.30 s filter on the modeled emission stands for the integration time, and the model has no clouded window. The self-bias is read electrically on the bias line, where sensors between the matching network and the electrode need no optical access and do not suffer from clouded windows [1].

What Moves It

The throttle valve on the pump line is clipped to 4 to 100 %, the source and bias generators to 0 to 2600 W and 0 to 700 W, and the backside helium to 0.5 to 25 Torr.

TECHNOLOGY Final Element: A Radio-Frequency Generator and Its Match

The source and bias powers come from radio-frequency generators, which typically have an output resistance of 50 Ω. Connected straight to a plasma, a generator would deliver its power poorly, so a matching network between generator and electrode presents the generator with a purely resistive load equal to its own. Its variable elements, often two vacuum capacitors called load and tune beside a fixed inductor, are adjusted automatically as the plasma impedance changes with pressure, temperature and gas flow. The network also protects the generator from reflected power. Setting one up takes real expertise, because every wire has appreciable inductance and capacitance at these frequencies [1].

The match efficiency of scenario 6 is the model's form of a detuned network. A tenth of the forward power reflects while the generator still reports 1200 W, so on a trend the forward power stays flat while the density falls.

Where the Controller Runs

The four loops and the sequencer belong to the etch tool, which runs the recipe: setpoints of pressure, power and gas flow for each step, developed by experiment for each product and distributed to the fabs [1]. Supervisory and lower-level loops run with cycle times from milliseconds to minutes, and what counts as real time has to be stated for each application [19]. The app scans the four loops every 20 ms and the predictive controller plans every 0.1 s, while a wafer spends about two minutes in the chamber. The structure of this chapter is three layers in that controller: four PI blocks in velocity form, a step sequencer that executes the recipe, and an endpoint block that ends one step.

> **IN PRACTICE** **A Production Recipe Stops One Step on Light**
>
> A production trench etch studied in 2011 ran five steps. A timed step removed about three quarters of the exposed nitride layer, a gentler chemistry removed the rest and stopped on the endpoint signal of a monochromator, and an over-etch of fixed length followed. The main trench step was timed, and operators set its length off-line, adjusting it by hand for drift. Chambers drift as etch products coat their walls, and the first one or two wafers of a lot often etch differently. Maintenance cleans the walls and the viewing window, which restores the endpoint signal to full strength [1].

What Is Modernizing It

Three developments bear on this chamber. Multivariate endpoint methods read the whole emission spectrum instead of one line, which keeps the endpoint usable as the open area shrinks [1, 8]. Virtual metrology uses the in-situ measurements of every wafer, the emission spectrum and the electrical signals of the RF line among them, with models that estimate the etch result before any metrology tool sees the wafer [1]. And a 2024 review of the field, which faces three-dimensional devices and dimensions approaching the atomic scale, names digital twins of etch chambers, built from hybrid models and machine learning, among its opportunities [20]. Real-time multivariable feedback on measured plasma variables, the regulatory layer of this chapter, was demonstrated on a reactive ion etcher in 1995 [17].

7.8 Limitations

The most important limitation is the one Section 7.6.3 ran into. Wafer properties are scored from the chamber state at the end of the etch rather than integrated over the etched depth, so a defect confined to the top or bottom of a feature does not appear on the scorecard even when the traces show the excursion that caused it. Any conclusion about transients inside the product must be read from the traces rather than from the yield.

The plasma is a global model with two lumped neutral species, no spatial structure and a normalized rather than absolute electron density. Real reactors are designed around their spatial structure, and coil geometry, gas injection and wafer edge effects are what a process engineer adjusts to correct uniformity. Equation (7.6) is a fitted shape that reproduces the interior optimum and cannot tell anyone where it is on a particular tool.

Aspect-ratio dependent etching is absent. Real features etch more slowly as they deepen, and the resulting dependence on feature size is the reason a recipe that works on one product does not work on the next. The application carries an open-area load effect but no feature-scale transport, so its sidewall angle is uniform with depth where a real profile is not.

Chamber memory is one scalar. Real chambers accumulate several distinguishable conditions, respond differently to each cleaning chemistry, and drift over weeks as well as over wafers, which is why production tools carry run-to-run control layers that adjust the recipe between wafers on

measured results [12]. Nothing in this application represents that layer, and it carries most of the advanced control in a production fab. Chapter 6 adds it to an etcher whose measurement arrives two lots late. The endpoint is also only a threshold. A moving horizon estimate or a Kalman filter driven by the emission signal and the clearing model of (7.9) would follow the cleared fraction continuously instead of declaring one crossing, and would size the over-etch from the estimated spread instead of a fixed percentage.

The predictive controller's model is linear about one operating point and leaves out what the plant does elsewhere. It knows nothing of ignition, so a strike is to the plan only the two generators going from zero to their commands. Its wafer element is the 15 s lag at 8 Torr of helium, while at the 0.5 Torr floor the plant's wafer gain to bias power is 0.844 °C/W with a 104.5 s time constant against the model's 0.121 and 15 s, because the backside conductance scales with $p_{\mathrm{He}}^{0.7}$. The throttle's effect on the pressure departs from the model by 7.8 % for a 1.4-point step, and the wall coverage, the match, the seal, the flow calibration and the pattern loading enter only through the output and heat-load disturbances of the estimator. The plant also has no model of reflected RF power, arcing or generator slew, so the harm the source rate limit prevents is not scored.

Finally, there is one wafer state rather than a wafer map, so the application reports a single non-uniformity number and cannot show where on the wafer the problem is, which is the first question a process engineer would ask.

7.9 Discussion

Three conclusions generalize beyond this chamber.

The first is that fast loops on a slow product convert a dynamic control problem into a static optimization problem, and the engineering effort should follow. Table 7.1 and (7.1) together say that the regulatory layer here is a solved problem, and Tables 7.4 and 7.5 say that the interesting decisions are all about where to put the setpoints. A course that spent its time on tuning alone would not prepare a student for any of the decisions that matter on this plant. The predictive controller of Table 7.8 is the measured form of that claim, because the same four loops retuned once recover 97 to 98 % of its gain on the wafer and its interaction terms are worth about a second per wafer.

The second is that a control element need not be a loop. The endpoint detector in Table 7.7 has no setpoint, no actuator and no error signal, and it is worth more than the four PI controllers combined because it is the only element that measures the product rather than the process. The general form is that a measurement taken on the product usually outperforms a measurement taken on the equipment making it, and endpoint detection is universal in this industry and rare elsewhere because here the product measurement is available while the process is running. Where it is not, Volume 8, Chapter 8 can end each coating step on an estimated thickness instead.

> ⚠ **OPEN CHALLENGE** **Finding the End When Almost Nothing Is Open**
>
> Endpoint detection is easy when more than a tenth of the wafer is open to the etch, because the emission changes clearly. As devices shrink the open area falls, and the change in the signal sinks toward the noise. Molecular etch gases give a continuous spectrum with no clean line to watch, and the etched mask adds an emission of its own [1]. Multivariate methods read the full spectrum: one in 2023 raised the signal-to-noise ratio 2.78 times over a single wavelength on wafers with 1.0 % of the area open [8]. Progress would be an endpoint that fires reliably at any open area. The detector of this chapter thresholds one modeled signal that falls cleanly as the film clears.

The third is that regulatory performance can be perfect while the plant produces nothing of value. Figure 7.2 shows the pressure, density and self-bias loops on setpoint to the displayed precision in the middle of a main etch and a yield of zero, which is the same conclusion the other applications in this series reach from different directions [2, 4].

7.10 Conclusion

A plasma etch chamber settles far faster than the wafer it is processing, its fastest loop within 0.15 s and its slowest within 45 s against a wafer that occupies the chamber for about two minutes, and its four natural loops are nearly decoupled despite a gain matrix in which the source generator moves all four measured variables. That makes it the one plant in this series where tuning is genuinely a secondary concern, and the application is built to demonstrate what takes its place. Ion energy buys rate, uniformity and profile and sells mask selectivity, at 42 % against 35 % across the useful range. Pressure has a uniformity optimum near 23 mTorr and a sidewall that keeps improving below it. And in a graded cassette after a wet clean, a timed open-loop recipe scraps all eight wafers, retuning the idle loops is worth nothing, closing them recovers none, and an optical endpoint detector that has no setpoint at all recovers seven of seven, because the main etch a fresh chamber actually needs varies by twelve percent from wafer to wafer and a timer cannot know it. A model predictive controller over all four actuators cuts the time the wafer spends outside its band after a recipe step from 37.1 to 3.5 s per wafer, but the same loops retuned once reach 4.1 s. Beyond about a second of wafer time per wafer, the gain that belongs to its interaction terms is at the strike, where the pressure moves 0.78 instead of 2.38 mTorr.

KEY POINTS

- The loops settle far faster than a wafer is etched, so the setpoints, not the tuning, decide the wafer.
- Ion energy buys rate, uniformity and profile and sells mask selectivity, 42 % against 35 %.
- After a wet clean a timed recipe scraps three wafers while the plasma loops hold their setpoints, because the wall polymer that sets the rate is not measured.
- An optical endpoint with no setpoint recovers seven wafers of seven in the graded cassette, and retuning or closing the loops recovers none.
- A predictive controller cuts the wafer’s time outside its band from 37.1 to 3.5 s, and the loops retuned once reach 4.1 s.

Where else this applies. A phase sequencer above loops in Volume 8, Chapter 7, a batch step ended on an estimate of the product in Volume 8, Chapter 8, and a triangular gain matrix in Volume 6, Chapter 2.

References

[1] S. Lynn, “Virtual metrology for plasma etch processes,” Ph.D. dissertation, National University of Ireland Maynooth, 2011. [Online]. Available: https://mural.maynoothuniversity.ie/id/eprint/2657/

[2] D. E. Seborg, T. F. Edgar, D. A. Mellichamp, and F. J. Doyle, III, *Process Dynamics and Control*, 4th ed. Hoboken, NJ: John Wiley & Sons, 2016.

[3] B. W. Bequette, *Process Control: Modeling, Design, and Simulation*. Upper Saddle River, NJ: Prentice Hall, 2003.

[4] K. J. Åström and R. M. Murray, *Feedback Systems: An Introduction for Scientists and Engineers*, 2nd ed. Princeton, NJ: Princeton University Press, 2021.

[5] M. A. Lieberman and A. J. Lichtenberg, *Principles of Plasma Discharges and Materials Processing*, 2nd ed. Hoboken, NJ: John Wiley & Sons, 2005.

[6] J. W. Coburn and H. F. Winters, “Ion- and electron-assisted gas-surface chemistry: An important effect in plasma etching,” *Journal of Applied Physics*, vol. 50, no. 5, pp. 3189–3196, 1979.

[7] K. Hirobe and T. Tsuchimoto, “End point detection in plasma etching by optical emission spectroscopy,” *Journal of The Electrochemical Society*, vol. 127, no. 1, pp. 234–235, 1980.

[8] S. Lee, H. Choi, J. Kim, and H. Chae, “Spectral clustering algorithm for real-time endpoint detection of silicon nitride plasma etching,” *Plasma Processes and Polymers*, vol. 20, no. 6, p. e2200238, 2023.

[9] J. O. Clevenger, M. J. Buie, and N. L. Sandlin, “Effect of chamber seasoning on the chrome dry etch process,” in *Proceedings of SPIE*, vol. 5130, 2003, p. 92.

[10] M. Surendra, C. R. Guarnieri, G. S. Selwyn, and M. Dalvie, “Across wafer etch rate uniformity in a high density plasma reactor: Experiment and modeling,” *Applied Physics Letters*, vol. 66, no. 18, pp. 2415–2417, 1995.

[11] S. Gaddam and M. W. Braun, "Etch chamber condition-based process control model for shallow trench isolation trench depth control," in *IEEE/SEMI Conference and Workshop on Advanced Semiconductor Manufacturing*, 2005, pp. 17–20.

[12] J. Moyne, E. del Castillo, and A. M. Hurwitz, Eds., *Run-to-Run Control in Semiconductor Manufacturing*. Boca Raton, FL: CRC Press, 2018.

[13] K. J. Åström and T. Hägglund, *Advanced PID Control*. Research Triangle Park, NC: ISA – The Instrumentation, Systems, and Automation Society, 2006.

[14] S. Skogestad, "Simple analytic rules for model reduction and PID controller tuning," *Journal of Process Control*, vol. 13, no. 4, pp. 291–309, 2003.

[15] S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Control Engineering Practice*, vol. 11, no. 7, pp. 733–764, 2003.

[16] E. F. Camacho and C. Bordons, *Model Predictive Control*, ser. Advanced Textbooks in Control and Signal Processing. London: Springer, 2007.

[17] B. A. Rashap, S. Lafortune, J. R. Moyne, D. Teneketzis, F. L. Terry, M. E. Elta, H. Etemad, J. P. Fournier, J. S. Freudenberg, M. D. Giles, J. W. Grizzle, P. T. Kabamba, and P. P. Khargonekar, "Control of semiconductor manufacturing equipment: Real-time feedback control of a reactive ion etcher," *IEEE Transactions on Semiconductor Manufacturing*, vol. 8, no. 3, pp. 286–297, 1995.

[18] K. R. Muske and T. A. Badgwell, "Disturbance modeling for offset-free linear model predictive control," *Journal of Process Control*, vol. 12, no. 5, pp. 617–632, 2002.

[19] K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, A. Hahn, S. Saravia, A. Sherule, and M. Thompson, "Guide to operational technology (OT) security," National Institute of Standards and Technology, Gaithersburg, MD, Special Publication 800-82r3, 2023.

[20] G. S. Oehrlein, S. M. Brandstadter, R. L. Bruce, J. P. Chang, J. C. DeMott, V. M. Donnelly *et al.*, "Future of plasma etching for microelectronics: Challenges and opportunities," *Journal of Vacuum Science & Technology B*, vol. 42, no. 4, p. 041501, 2024.

7.11 Exercises

7.1 Concept check · 2 min

In scenario 7 the pressure, density and self-bias loops read 20.000 mTorr, 1.0000 and 120.00 V at the end of every main etch, yet the timed recipe scraps wafers 1 to 3 and passes wafer 4. What changed between wafer 1 and wafer 4?

(a) The wafer temperature loop settled, so the chemical term of the etch rate rose

(b) The density loop wound up its integral, so the plasma density rose during the etch

(c) The emission detector learned the clearing time and lengthened the main etch

(d) The wall polymer coverage rose, so fewer radicals were lost to the walls

7.2 Concept check · 1 min

Under the open-loop recipe of scenario 6 the RF match drifts so that a tenth of the forward source power reflects. The bias generator still delivers 150 W. What does the self-bias do?

(a) It rises, because a thinner plasma has fewer ions to carry the bias current

(b) It falls, because the plasma receives less power and develops less voltage

(c) It stays at 120 V, because the bias generator is a separate supply

(d) It falls, because less dissociation lowers the chamber pressure

7.3 Concept check · 2 min

In the graded challenge the emission endpoint alone is worth seven good wafers of seven. Which description of the endpoint detector is correct?

(a) A PI loop that trims the source power to hold the emission at its 0.55 setpoint

(b) A threshold on a filtered emission signal that ends the main etch, with no setpoint or actuator

(c) A model that predicts each wafer's clearing time from the chamber rate before the etch

(d) A derivative trigger on the emission signal, armed after a minimum main etch time

7.4 Theory and calculation · 7 min

Use Equation (7.5) and the gain matrix of Table 7.2. The design point is 120 V of self-bias at 150 W of bias power and a density of 1.000.

(a) Find $\partial V_{dc}/\partial P_{\text{bias}}$ and $\partial V_{dc}/\partial n_e$ at the design point. With the density gain of 8.94×10^{-4}/W from the table, predict the source-to-V_{dc} entry to three significant figures. *Checkpoint:* 0.400 V/W; –60 V per unit density; –0.0536 V/W.

(b) The density and pressure rows with the source and throttle columns form a 2×2 block. Compute its relative gain $\lambda = 1/(1 - K_{12}K_{21}/(K_{11}K_{22}))$ for density on the source and pressure on the throttle, to three decimals. *Checkpoint:* 1.073.

(c) Why are the self-bias and wafer-temperature diagonal entries of Equation (7.1) exactly 1, and what does the relative gain of density on the throttle say about that pairing? *Checkpoint:* The bias and helium columns reach no row above their own, so the matrix is block triangular; $1 - 1.073 = -0.073$ is negative, so that pairing is rejected.

7.5 Theory and calculation · 7 min

The wafer obeys $C_w\dot{T}_w = \eta\,(P_{\text{bias}} + s\,P_{\text{src}}) - h(p_{\text{He}})\,(T_w - T_{\text{chuck}})$ with $h = h_0\,(p_{\text{He}}/8\,\text{Torr})^{0.7}$, $C_w = 60\,\text{J/K}$, $h_0 = 4.00\,\text{W/K}$ and $T_{\text{chuck}} = 20\,^\circ\text{C}$. At the design point the wafer holds 60 °C at 8 Torr.

(a) From the wafer row of Table 7.2, find the fraction η of bias power that reaches the wafer and the source share s, to three decimals. *Checkpoint:* $\eta = 0.484$, $s = 0.150$.

(b) Predict the helium entry of that row from the balance. *Checkpoint:* –3.50 °C/Torr.

(c) A cold wafer warms with the helium at its 0.5 Torr floor. Find the conductance, the time constant C_w/h and the steady gain of the wafer to bias power there, and compare them with the predictive controller's 15 s and 0.121 °C/W. *Checkpoint:* 0.574 W/K, 104.5 s, 0.843 °C/W; about seven times both.

7.6 Theory and calculation · 7 min

Use Equation (7.9) for a design wafer: film 500 nm, non-uniformity 2.2 %, etch rate 5.831 nm/s and an 86 s main etch. The emission signal is $(1 - c) + 0.12$ and the detector fires when the filtered signal falls below 0.55, which the filter lag delays to 64 % cleared.

(a) Find the cleared fraction at the threshold, the width of the clearing transition, and the depth past the film thickness at which the detector fires. *Checkpoint:* 0.57; 6.6 nm; 3.80 nm.

(b) Find the further depth needed to clear 0.995 of the wafer, the time it takes, and that time as a percentage of the main etch. *Checkpoint:* 31.1 nm, 5.34 s, 6.2 %.

(c) Find the cleared fraction after a 5 % over-etch, and the over-etch, as a percentage of the same 86 s main etch, that a wafer of 4 % non-uniformity would need if the detector still fired at 64 % cleared. *Checkpoint:* 0.988; 11.3 %.

7.7 App challenge · 8 min

Load scenario 10, How much over-etch is enough (default seed), and run it as loaded at 5 s/s until the Yield tile reads 3 / 6 (wafer 6 is graded at about 12.5 min; the next wafer is graded after 15 min). Pause and export the CSV. Its time column is Time (min); the wafer number is `wafer_no`, and a wafer's rows run from its strike to the next.

(a) For each wafer, find the largest `cleared` and the largest `under_nm` in its rows, to ±0.001 and ±0.1 nm, and the over-etch setting `ctrl_over_pct` in its rows whose `phase` is `over`. Which wafers are good, and why do the others fail? *Checkpoint:* Wafers 1–2: 5 %, 0.987, 1.15 nm, not cleared. Wafers 3–5: 30 %, 1.000, 6.16 nm, good. Wafer 6: 75 %, 1.000, 15.18 nm, punch-through.

(b) The over-etch of the 30 % wafers lasted 25.8 s and that of wafer 6 64.5 s. Assume the loss grows in proportion to the over-etch time once the film has cleared. Find the loss per second, and the largest over-etch percentage of the 85.96 s main etch that keeps the loss under the 12 nm limit, to the nearest percent. *Checkpoint:* About 0.23 nm/s; about 59 %.

7.8 App challenge · 8 min

Load scenario 11, One plan for four loops: MPC against the closed loop (default seed), and run it as loaded at 5 s/s until the clock passes 10.0 min. Pause and export the CSV (its time column is Time (min)). Press Reset, set Chamber control to Closed loop, run again to 10.0 min and export a second file.

(a) In each file, find the largest `Tw_C`, to ±0.1 °C. *Checkpoint:* 62.0 °C under the predictive controller, 67.5 °C under the loops.

(b) In each file, read `Tw_C` and `he_cmd_Torr` in the row at 310 s (Time (min) 5.16666667), ten seconds after the self-bias setpoint steps to 150 V, to ±0.02. *Checkpoint:* 60.37 °C and 10.99 Torr; 64.05 °C and 9.84 Torr.

(c) Explain from the helium command why the plan's wafer stays inside its 3 °C band through the step while the loops' wafer leaves it. *Checkpoint:* The plan's model carries the bias-to-wafer gain, so it raises the helium as the bias power rises; the loop waits for the temperature error and is tuned slowly.

7.9 App challenge · 9 min

Load the graded challenge, Challenge · A cassette after a wet clean (default seed). Before Run, set Main etch ends on to Emission and change nothing else. Run at 5 s/s and Pause once the clock passes 15.0 min and before 16.5 min. Then reload the scenario with the reload button beside the scenario list, set only Over-etch (under Recipe timing) to 32 %, and repeat. The target is at most one scrapped wafer and at least five good ones after 900 s.

(a) Read the Yield tile after each run and say whether the target is met. *Checkpoint:* 7 / 7, met; 5 / 7, not met.

(b) Which wafers does the second run scrap, as the wafer notes say, and why can more over-etch not save them? *Checkpoint:* Wafers 5 and 6, not cleared. The more open product from 450 s etches slower, and a percentage of a fixed 86 s main etch is still a fixed time.

(c) Closing the four loops on their own leaves every wafer scrapped. Why does the endpoint, which has no setpoint and no actuator, meet the target alone? *Checkpoint:* It measures the product. The rate falls with the clean, the match and the product change, and the endpoint lengthens each main etch to match.

7.10 Application area · 3 min

Scenario 5 deletes the 8 s stabilization step to gain 7 % more wafers an hour, and every wafer is still graded good.

(a) What fraction of the 500 nm film is etched before the pressure is back within 0.6 mTorr of setpoint, and why does the Yield tile not see it? *Checkpoint:* About 1.1 % (5.6 nm); the wafer is scored from the chamber state at the end of the etch.

(b) What would a process engineer check before accepting the change? *Checkpoint:* The top of the feature profile on sampled wafers, and the strike transient in the tool's own trace data.

7.11 Application area · 4 min

After a wet clean, a fab running the timed recipe of scenario 7 burns seasoning wafers before product.

(a) From Table 7.6, how many seasoning wafers does it need, and how much tool time do they take at the scenario's 110.4 s per wafer? *Checkpoint:* Three; about 331 s, or 5.5 min.

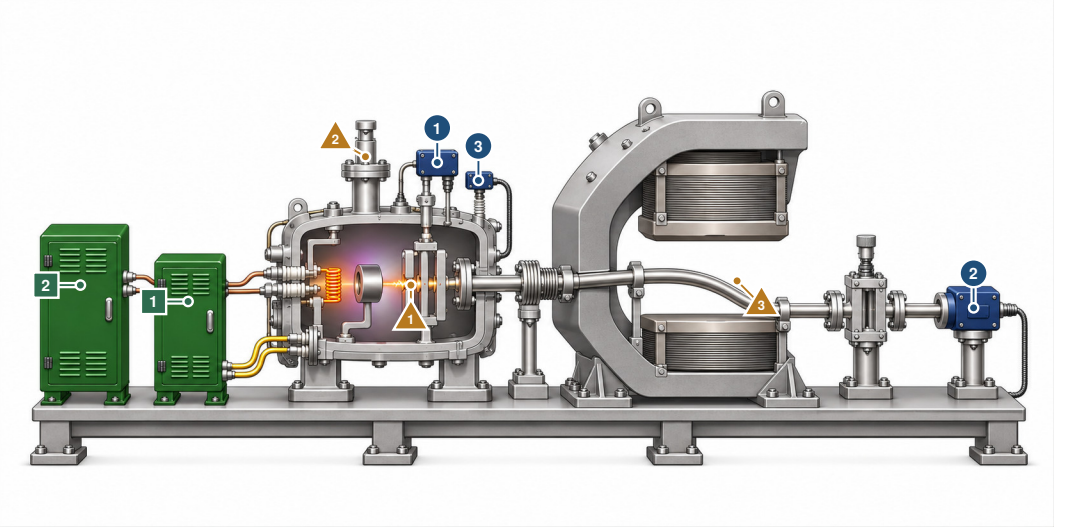
(b) A run-to-run controller lengthens the main etch from the film left on each wafer, measured before the next one starts. Which wafers can it save, and what can the emission endpoint do that it cannot? *Checkpoint:* Wafers 2 onward at best; wafer 1 is lost. The endpoint saves wafer 1 too, because it measures the wafer being etched.

7.12 Review the agent · 5 min

An AI assistant reviews scenario 11 and writes: "The predictive controller cuts the time the wafer spends outside its band from 37.1 to 3.5 s per wafer and the hottest wafer from 67.5 to 62.0 °C. The relative gain array is within 0.073 of the identity, so the four loops cannot be the weak point, and the gain must come from the bias-to-wafer term the array cannot see. Replace the four loops with the predictive controller." Find the flaw, and say how you would check it in the app.

CHAPTER 8

An Ion Source Simulator for Cascade Control and Recovery From a Measurement That Stops Meaning Anything



The controller moves 1 bias supply, heats the cathode 2 filament supply

The controller reads 1 extraction current 2 selected beam (Faraday cup) 3 gap voltage and supply current

What disturbs it 1 extraction spark 2 gas flow below its reading 3 beam lost in the magnet

Notice the spark at the extraction gap: for a few milliseconds the beam readings collapse while the hot filament and cathode beside it do not change. Illustration, not to scale.

IN PLAIN TERMS **When a Reading Stops Telling the Truth**

The main idea. Chip makers push boron atoms into silicon wafers with a narrow beam of charged atoms. A hot source makes the beam, and automatic controls keep it steady by turning the source's heating power up or down.

The challenge. Now and then the high voltage that pulls out the beam sparks, and the beam reading drops for a moment while the source is fine. A control that trusts the reading turns the heat up, and the beam overshoots when the voltage returns.

Why it matters. Holding the controls still through the spark, then putting the missed boron back only where it was missed, keeps the wafer good.

THREE WAYS TO READ THIS CHAPTER

Explore	12 min	The first page, the history boxes of Sections 8.1 and 8.2, the fact box of Section 8.3, the try-it and fact boxes of Section 8.6, and the key points
Study	45 min	The whole chapter with its pause points and the twelve exercises
Practice	15 min	The first page, Section 8.6, Section 8.7, Section 8.8 and its open challenge

The control question A spark across the extraction gap halves the beam voltage in half a millisecond while the cathode, with a time constant of seconds, notices nothing. What should a cascade on the beam do with a measurement that has stopped meaning anything, and what does each choice cost in source heat, beam overshoot and missed dose?

Learning objectives

- Name the current each loop of the three-level cascade reads and moves, and explain why the fast loop moves bias current rather than filament current.
- Explain why back-calculation anti-windup does not stop a loop from chasing a spark, from the heat the cathode stores.
- Compute how long supply-current surge and gap collapse coincide, and choose a detection window and a measurement gate from it.
- Compare no repaint, a whole-pass repaint and a local repaint driven by a per-strip dose ledger.
- Judge when a retry limit should turn automatic restarts into a trip.

Prerequisites Sections 1.6.2 (velocity form, anti-windup, bumpless transfer and loop action), 1.4.2 (measurements that stop meaning anything), 1.6.1 (control period, speed and seed) and 1.6.4 (scenarios, scorecard and CSV).

The app https://apmonitor.com/apps/ion-implant-source/

See also Volume 2, Chapter 2 for the velocity-form PID with back-calculation, and Volume 2, Chapter 3 for windup at a true actuator limit. Volume 3, Chapter 2 for matching loop structure to where a disturbance enters, and Volume 3, Chapter 3 for the speed an inner loop must reach. Volume 8, Chapter 13 for a product history kept apart from the instruments that act on it.

8.1 Introduction

Every feedback loop carries an assumption that is almost never written down, which is that the measurement still represents the plant variable the loop is meant to regulate. Saturation does not violate that assumption and neither does noise, and the standard remedies for both leave it intact. A protection event violates it completely while leaving every quantity the controller can see numerically valid.

The clearest place to see it is an ion implanter. A tungsten filament at 2600 K heats a separate cathode at 2200 K by electron bombardment, the cathode fills an arc chamber with boron trifluoride ions, an extraction electrode at 10 kV pulls them out, and an analyzer magnet keeps one mass in five. When that extraction gap sparks, the supply current goes to its limit, the gap voltage halves in about half a millisecond, and the beam goes with it. The two thermal masses notice nothing. A proportional-integral controller on extraction current then reads a thirty-five milliamp error in a variable it is responsible for and does the only thing it knows, which is to raise source power. Measured on the simulator of this chapter, that drives the bias command from 2.493 A to its 4.0 A limit, adds 28.2 K to the cathode, and overshoots the beam by 14.75 % when high voltage returns.

> ⧖ HISTORY **From Curiosity to Production Doping (1950s to 1978)**
>
> Ion beam doping of semiconductors was developed from the early 1950s, at first as an experimental curiosity of solid-state physics. Early workers were drawn to the damage that ions do to a crystal and overlooked the real prize: dopant atoms placed with control and then activated by annealing above 800 °C. Device engineers with manufacturing problems to solve brought it into transistor production, where it gave good control of where a junction sits and let threshold voltages be tailored [1]. By 1978 implanters are considered to have come of age [2]. Some forty years into its development the industry was selling about two billion dollars of implanters a year and had installed thousands of machines [3].

Back-calculation anti-windup does not prevent any of it, and the reason is worth stating precisely. Back-calculation answers the question of whether the actuator has run out of room [4, 5]. During the

dropout it has, for 32 ms, and back-calculation duly holds the integral below 0.23 A. The proportional term alone already asks for the limit, and the cathode stores the heat. The actuator is being spent on an error that does not exist, and no bound on the integrator changes that. The remedy belongs one layer up, in a measurement-quality gate that invalidates the signal on corroborating evidence, an explicit output hold, a frozen rather than merely limited integrator, and an integral state reconstructed from the held output before the loop is handed back [6, 7].

The same structure appears wherever a protection system takes a process temporarily outside its controller's domain. A compressor on anti-surge recycle (Volume 4, Chapter 2), a fired heater in a burner management sequence, a reactor after a trip and a converter in a fault-current limit all produce measurements that keep arriving and stop being evidence. The implanter is the cleanest teaching case because the separation of time scales is extreme. The fault is five milliseconds long and the plant it appears to describe has a four second time constant, so there is no way to confuse the two.

There is a second lesson underneath. Four quantities in an implanter are called current and no two of them are the same thing. At the nominal recipe the source discharge runs at 10.0 A, the extraction supply delivers 35.00 mA, and 7.918 mA of mass-selected $^{11}B^{+}$ reaches the wafer. A twelve percent analyzer detune moves the third by −12.00 % and leaves the second unchanged. One second later the outer loop has raised the second by 0.25 % while the third is still 11.76 % low, so an outer loop reading only the Faraday cup drives the source harder to fix a beamline problem.

This chapter makes measurement-quality gating measurable on an interactive ion source. With the same gains and the same spark the fault-chasing index is 7.46×10^{-2} A s without the gate and 8.28×10^{-3} A s with it, and a local dose repaint returns a wafer to 0.21 % strip non-uniformity and +0.21 % dose where a whole-pass repaint overdoses it by 16.58 %. The chapter follows three threads. The first is a sixteen-state reduced source and beamline model with a switched-conductance high-voltage fault, a coincidence detector at the physics step and a three-level cascade, written closely enough to rebuild. The second separates two judgments that are usually made together, the latching of a fault and the invalidation of one measurement, and measures the timing cost of conflating them in actuator movement. The third is a scenario pair in which one plant, one seed and one spark produce 14.75 % and 0.29 % of beam overshoot under two architectures that differ by one decision. Anti-windup, cascade control and the Child-Langmuir law are all standard. What is particular to the implanter is how they combine.

Section 8.2 places the problem in its domain. Section 8.3 gives the source and beamline model, and Section 8.4 the three loops, the detector and the supervisor as implemented. Section 8.5 separates the fault latch from the measurement gate, Section 8.6 reports the measured outcomes of the twelve scenarios, Section 8.7 names the real sensors, supplies and controller platform, and Section 8.8 states what the model leaves out. The Exercises at the end of the chapter use the scenarios.

8.2 Sources, Sparks and Held Outputs

The ion sources used in implantation descend from the isotope separators of the 1960s [8], and the History box below tells how the first filament source came to dope silicon. Its weakness is that the filament is the plasma electrode, so filament erosion and plasma conditions are the same variable.

The indirectly heated cathode architecture separates them. A filament behind a shield is heated by its own current, its electrons are accelerated across a few hundred volts onto a separate cathode, and that bombardment heats the cathode to emission temperature. The cathode is then the plasma electrode and the filament is a heater, which gives longer life and a cleaner control handle [9]. Langmuir probe measurements inside such a source relate electron temperature and plasma density to the source control parameters that an operator actually sets [10], and the fluorine chemistry of a boron trifluoride feed both feeds the process and attacks the hardware, which is why source lifetime is a live engineering subject rather than a settled one [11].

HISTORY **From Isotope Separators to Chip Doping (1963 to 1992)**

In 1963 a new ion source was described for electromagnetic isotope separators [8], machines that sort the ions of an element by mass to purify one isotope [12]. Its heated filament, a straight tungsten wire placed parallel and close to the extraction slit, drives the arc discharge, and the wear of that wire by evaporation and sputtering limits the life of the source [12]. Thirty years later it was still the most widely used ion source. Its sharply focused milliampere beams of most elements had made it a source of choice for the production doping of silicon devices, and filament lifetime was still a subject of study [13].

The physics that limits extraction is older than either source. The three-halves power dependence of space-charge-limited current on accelerating voltage was established for thermionic diodes in the 1910s [14, 15], and the temperature dependence of thermionic emission itself was settled in the same decade [16]. Both appear in this model in normalized form, because a control study needs the exponent and the bound rather than the absolute current.

On the control side the ingredients are old and the combination is not. Integrator windup and the back-calculation remedy were given their standard treatment in 1989 [4] and placed in a unified framework five years later [5]. Anti-windup, bumpless transfer and reliable switching were later shown to be one problem viewed three ways [6], which is exactly the observation this application is built on. Fault-tolerant control supplies the vocabulary for what happens when a signal is present but no longer informative [7]. What the implanter literature adds is the consequence on the product, because an implant specification is an integral rather than a level and an interruption therefore leaves a spatial defect. Scanned beam uniformity is controlled by the scan itself rather than by the source [17], and a fault handler that did not record where the scanner was cannot repair what the fault cost [18].

8.3 Process Model

Boron trifluoride enters arc chamber AC-101 through mass flow controller FC-101. Tungsten filament FIL-101 is heated by supply PS-F at current I_f. Filament electrons are accelerated onto cathode CA-101 across bias supply PS-B at V_b and current I_b, and the bombardment power $V_b I_b$ is what heats the cathode. Cathode electrons are accelerated into the gas across the source arc potential V_a and confined by axial field B, and repeller RE-101 returns them through the chamber. Extraction electrode EX-101 at V_x and suppression electrode SU-101 at V_s form the beam, analyzer magnet AM-201 bends it, resolving slit RS-201 cuts it, and the scanned wafer is the load. The equipment and every symbol in the equations are drawn in Figure 8.1.

DID YOU KNOW? **Hundreds of Hours on One Cathode**

In 1998 an indirectly heated cathode source ran several times longer than sources with a bare filament cathode. Its cathode lasted from 70 hours at the highest discharge power to over 500 hours in moderate operation, and near 1 kW of discharge power its ions were mostly atomic rather than molecular [9].

Table 8.1 collects the three controlled variables, the three manipulated variables and what acts against them. Time is carried in seconds, currents in amperes for the supplies and milliamperes for the beam, temperatures in kelvin, electrode voltages in kilovolts and electron temperature in electronvolts. Sixteen states are integrated at $\Delta t = 0.2\,\mathrm{ms}$, the fault detection and the dose ledger run at that same step, and the three controllers run at 1 ms, 20 ms and 100 ms.

8.3.1 Two Thermal Masses in Series

The filament heats itself. Its resistive drive is quadratic in current and scaled by a drift factor r_f that represents thinning with age, and the cathode is heated by the bombardment power delivered across the bias potential with a small conduction term from the filament, $k_{fc} = 0.12$, giving Equations (8.1)

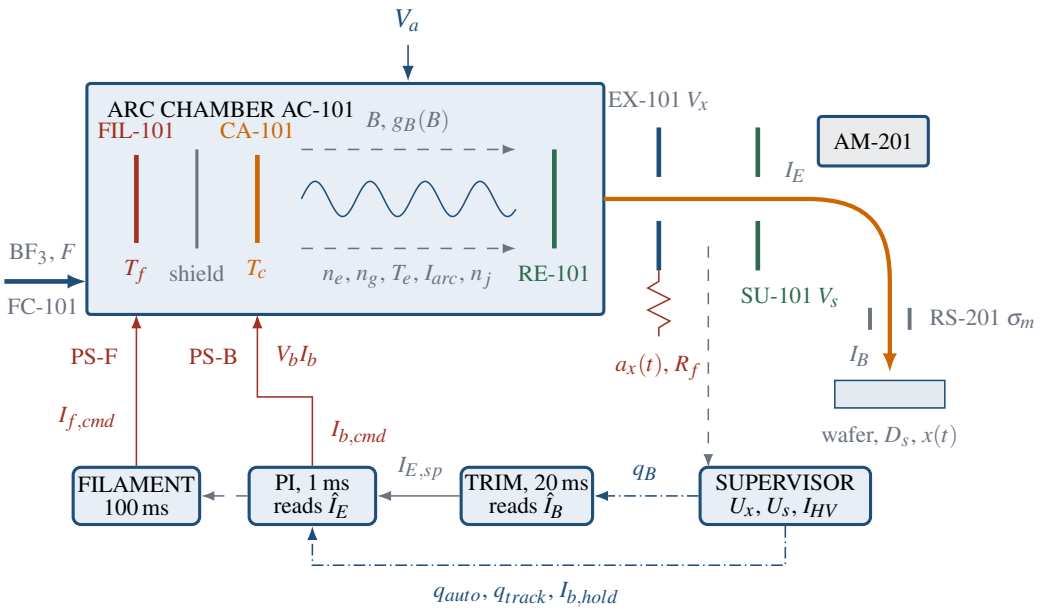


Figure 8.1 The indirectly heated cathode source, the beamline and the control path. Bombardment power V_bI_b heats the cathode, the cathode fills the chamber, and extraction pulls a mixed beam that the analyzer magnet and the resolving slit reduce to one mass. The zigzag at the extraction electrode is the switched conductance $a_x(t)$ through fault resistance R_f that constitutes a spark. The inner loop reads extraction current and sets the bias supply, the trim reads the selected beam and sets the inner setpoint, and the filament supervisor reads the bias command and sets the filament supply. The supervisor latches faults on the high-voltage evidence alone, reads the extraction current only to qualify a recovery, and owns the permissives that decide whether the inner loop and the trim may act on their measurements. In the control path, supply commands are drawn solid, signals read by a controller dashed and permissives dash-dotted.

and (8.2).

$$\tau_f \dot{T}_f = T_w + r_f \left(\frac{I_f}{I_{f,0}} \right)^2 (T_{f,0} - T_w) - T_f \tag{8.1}$$

$$\tau_c \dot{T}_c = T_w + \frac{V_b I_b}{V_b I_{b,0}} \left[1 + k_{fc} \left(\frac{T_f}{T_{f,0}} - 1 \right) \right] (T_{c,0} - T_w) - T_c \tag{8.2}$$

Both are written so that the nominal recipe is a fixed point, which makes $T_{f,0}$ and $T_{c,0}$ operating points rather than fitted constants. Neither state can move measurably in the five milliseconds a spark lasts, and that single fact is what the whole fault-handling design rests on.

8.3.2 Emission, Written as a Ratio

Direct evaluation of Richardson-Dushman emission at two thousand kelvin is numerically stiff and the control problem only ever needs the ratio to the reference, so the model carries the normalized form of Equation (8.3) with $\theta = T_c / T_{c,0}$.

$$\tilde{I}_e = \mathrm{clip}\left[\theta^2 \exp\left(\beta_R \left(1 - \frac{1}{\theta} \right) - \gamma_w w \right), 0, I_{e,max} \right] \tag{8.3}$$

The exponent β_R stands for $\phi / k_B T_{c,0}$ and is an adjustable parameter rather than a measured work function. At the shipped value of 9 the emission is 0.562 at 2090 K, 1.000 at 2200 K, 1.692 at 2310 K

Table 8.1 Controlled, manipulated and disturbance variables

Symbol	Role	Design value	Range
I_E	CV, inner, 1 ms	35.00 mA	10–80 mA
I_B	CV, outer, 20 ms	7.918 mA	1–12 mA
I_b	CV, supervisor, 100 ms	2.500 A	0–4.0 A
$I_{b,cmd}$	MV, inner loop	2.500 A	0–4.0 A
$I_{E,sp}$	MV, outer loop	35.0 mA	10–80 mA
$I_{f,cmd}$	MV, supervisor	155 A	100–190 A
T_f	state, unmeasured	2600 K	—
T_c	state, measured	2200 K	trip 2560 K
n_g	state, unmeasured	1.000	normalized
n_e	state, unmeasured	1.000	normalized
I_{arc}	state, measured	10.0 A	normalized $\times 10$
n_j	state, five species	mix of T_e	normalized
U_x	state, measured	10.0 kV	2–30 kV
U_s	state, measured	2.0 kV	0.5–5 kV
w	state, unmeasured	0 clean	0–1
E_{fault}	state, ledger	0 J	—
g_F	DV, unmeasured	1.000	0.80–1.20
η_{tr}	DV, unmeasured	1.000	0.70–1.05
a_x, a_s	DV, switched	0	$\{0,1\}$
D_s	product, 32 strips	1.000 at end	—

and 2.742 at 2420 K. Five percent of temperature is sixty-nine percent of emission, which is a plant whose gain cannot be quoted without also saying where it is standing. The deposit state w raises the effective work function as the source ages.

The bias supply is limited by what the filament can actually emit, $I_{b,act} = \min(I_{b,cmd}, I_{em,f})$, and both supplies are first order lags of time constants τ_{Ib} and τ_{If} toward their commands.

8.3.3 Gas, Electrons and the Intended Discharge

The normalized gas density follows the indicated flow through a calibration factor, the electron density follows the cathode emission through a bounded confinement factor, and the intended source discharge current is a first order proxy driven by all of them.

$$\tau_g \dot{n}_g = \frac{g_F F}{F_0} - n_g, \qquad \tau_e \dot{n}_e = \tilde{I}_e\, g_B(B) - n_e \tag{8.4}$$

$$\begin{aligned} g_B(B) &= \frac{B/B_0}{1 + B/B_0} \Big/ \frac{1}{2} \\ \tau_{arc}\, \dot{I}_{arc} &= \frac{V_a}{V_{a,0}}\, n_e n_g\, g_B(B) - I_{arc} \end{aligned} \tag{8.5}$$

The confinement factor of Equation (8.5) is normalized to unity at the nominal field and saturates by construction, because more field always raises the number of ionizing collisions and never without limit. The axial field bends the electrons into helices and the repeller returns them, and the model represents that as a bounded scalar rather than by solving trajectories.

8.3.4 One Gas, Five Ions

Five lumped species are carried, BF_3^+, BF_2^+, BF^+, B^+ and F^+. Each relaxes on a common time constant toward its share of the ionization rate $n_e n_g$, and the share is set by an effective yield and a

reduced rate coefficient.

$$\tau_i \dot{n}_j = f_j^*(T_e)\, n_e n_g - n_j, \qquad k_j(T_e) = \sqrt{T_e}\, e^{-E_j/T_e}$$
$$f_j^*(T_e) = \frac{y_j b_j k_j(T_e)}{\sum_\ell y_\ell b_\ell k_\ell(T_e)} \tag{8.6}$$

The yields y_j are 0.55, 1.00, 0.45, 11.5 and 0.35 in the order listed, and b_j is one except for B^+, where it is the B^+-fraction disturbance. The thresholds E_j are the gas-phase appearance energies rounded to a tenth of an electronvolt, 15.7 eV for BF_3^+, 15.9 eV for BF_2^+, 24.5 eV for BF^+, 31.0 eV for B^+ and 17.4 eV for F^+. They order the onsets and are not a claim about cross sections. The effective electron temperature is a bounded function of the arc potential, the gas loading and the deposits, with $k_V = 0.035\,\mathrm{eV/V}$, $k_p = 0.9\,\mathrm{eV}$ and $k_w = 1.8\,\mathrm{eV}$,

$$T_e = \mathrm{clip}[T_{e,0} + k_V \Delta V_a - k_p(n_g - 1) - k_w w,\ 2,\ 12] \tag{8.7}$$

in electronvolts, with $\Delta V_a = V_a - V_{a,0}$. At the nominal recipe $T_e = 6.00\,\mathrm{eV}$ and the mix is 19.8 % BF_3^+, 34.8 % BF_2^+, 3.7 % BF^+, 32.3 % B^+ and 9.5 % F^+. Raising the arc potential to 75 V takes T_e to 6.88 eV and the mix to 17.4, 30.7, 4.0, 39.3 and 8.6 %, so the parent falls and B^+ rises fastest, as the thresholds require.

8.3.5 Extraction and Mass Selection

All species share one aperture, so the extracted ion current is the smaller of what the plasma supplies and what the gap will pass.

$$I_{pl} = K_{pl} \sum_j n_j \sqrt{\frac{T_e}{T_{e,0}}}, \qquad f_j = \frac{n_j}{\sum_\ell n_\ell} \tag{8.8}$$

$$I_{CL} = 1.45\, I_{E,0} \left(\frac{U_x}{V_{x,0}}\right)^{3/2} \sqrt{\frac{m_{ref}}{m_{eff}}}, \quad m_{eff}^{-\frac{1}{2}} = \sum_j \frac{f_j}{\sqrt{m_j}} \tag{8.9}$$

$$I_{ion} = \min\left(I_{pl}, I_{CL}\right), \qquad I_E = I_{ion} + I_{par}\frac{U_x}{V_{x,0}} \tag{8.10}$$

The masses m_j are those of the ^{11}B ions, 68, 49, 30, 11 and 19 u, and $m_{ref} = 30\,\mathrm{u}$. The gain K_{pl} is calibrated so that the nominal recipe produces exactly the declared extraction current. At that point $m_{eff} = 25.13\,\mathrm{u}$ and the plasma supplies 33.20 mA against a space-charge limit of 55.45 mA, so extraction is plasma limited and raising the extraction voltage buys essentially no extra current while changing the implant energy. That inequality is the whole argument for recipe-locking V_x and manipulating I_b instead.

Beam optics collapse with the extraction field, and below about a third of the tuned voltage nothing useful is transported, $\kappa = \mathrm{clip}[(U_x/V_{x,0} - 0.30)/0.45, 0, 1]$. Each boron-bearing species splits into an ^{11}B and an ^{10}B component in the ratio f_{11}, the resolving slit passes a Gaussian window in mass-to-charge ratio, and the selected beam is Equation (8.11).

$$S_{j\nu} = \exp\left[-\frac{(m_{j\nu}/Z_j - \mu_{set})^2}{2\sigma_m^2}\right]$$
$$I_B = \kappa\, \eta_{tr} \sum_{j,\nu} S_{j\nu} f_\nu f_j I_{ion} \tag{8.11}$$

At the nominal recipe with $\mu_{set} = 11$ this gives 7.918 mA, which is 22.6 % of the extraction current.

8.3.6 Deposits and the High-Voltage Fault

A single lumped state represents conductive and corrosive deposits, $\dot{w} = k_{dep} n_g s - k_{clean} w$ with s the deposit-rate slider. It raises the effective work function in Equation (8.3) and it raises a stochastic glitch hazard, $\lambda_g = \lambda_0 \exp[\beta_w w + \beta_V (U_x/V_{x,0} - 1)]$. That hazard exists to make faults more likely in degraded scenarios and is never a predicted time to arc.

Each electrode is a first order supply whose time constant is switched by what is happening to it, Equation (8.12).

$$\dot{U}_k = \frac{U_{k,sp}^{eff} - U_k}{\tau_k}, \qquad \tau_k = \begin{cases} \tau_{coll} & a_k = 1 \\ 3\tau_{coll} & \text{HV inhibited} \\ t_r/3 & \text{recovering} \\ \tau_{HV} & \text{otherwise} \end{cases} \tag{8.12}$$

for $k \in \{x, s\}$, with $U_{k,sp}^{eff}$ zero while a fault is conducting or high voltage is inhibited. The supply current is the beam load plus the fault current, limited by the supply,

$$\begin{aligned} I_{HV} &= \min\left[I_{max},\; I_E + \sum_k a_k \frac{U_k}{R_f} \right] \\ \dot{E}_{fault} &= \sum_k a_k \frac{U_k^2}{R_f} \end{aligned} \tag{8.13}$$

which produces the signature the detector is written around. The supply current goes to its limit and the gap voltage collapses, both within a millisecond, while nothing thermal moves at all.

8.3.7 Scan and Dose

The wafer is thirty-two strips crossed by a scan at position $x(t)$ advancing at $1/t_{scan}$ per pass. Two ledgers are kept per strip. The committed dose X_s accrues whether or not the beam is on, and the delivered dose D_s accrues only when it is, Equation (8.14).

$$\dot{X}_s = \frac{\delta_s N_{st}}{n_{pass} t_{scan}}, \qquad \dot{D}_s = b_w(t) \frac{I_B}{I_{B,sp}} \frac{\delta_s N_{st}}{n_{pass} t_{scan}} \tag{8.14}$$

where δ_s is one while the scan is over strip s and zero otherwise, and b_w is the physical beam-enabled indicator, zero while high voltage is inhibited or below 35 % of its setpoint and, during a local repaint, over strips already at their committed dose. The deficit $\max(0, X_s - D_s)$ is what a repaint has to restore, and keeping the two ledgers separate is what makes the difference between putting dose back where it went missing and adding a pass everywhere. The committed ledger does not accrue during a repaint, because a repaint pass delivers dose the ledger already owed. The delivered ledger follows the simulated true beam and the physical beam-enabled indicator and deliberately not the controller's measurement-valid flag, so a dead instrument never erases dose that was actually delivered.

8.3.8 Initialization and Numerics

A warm scenario initializes at the nominal recipe, $T_f = 2600\,\mathrm{K}$, $T_c = 2200\,\mathrm{K}$, $I_f = 155\,\mathrm{A}$, $I_b = 2.5\,\mathrm{A}$, $n_g = n_e = 1$, n_j at the mix of $T_e = 6\,\mathrm{eV}$, $U_x = 10\,\mathrm{kV}$ and $U_s = 2\,\mathrm{kV}$. A cold scenario sets every one of those to zero or to the wall temperature and the supervisor to OFF. Integration is fixed-step fourth order Runge-Kutta at 0.2 ms. The fast task runs after every integration step and owns the beamline evaluation, the fault model, the coincidence detector, the supervisor transitions and the dose ledger, because hardware protection has to act before any supervisory controller can. The historian logs at 100 Hz, which is coarse enough that the sub-millisecond shape of a fault is not on the strip charts and fine enough that a fifteen millisecond event always lands on at least one sample. The application carries a separate fast-event display for exactly that reason. It keeps the last 1250 integration steps, the 250 ms its axis shows.

Table 8.2 Model parameters. All are adjustable in the model dialog unless marked hidden or fixed.

Symbol	Meaning	Value	Unit
$I_{f,0}$	Nominal filament current	155	A
$I_{b,0}$	Nominal bias current	2.5	A
V_b	Bias voltage	350	V
V_a	Source arc voltage	50	V
B	Source magnetic field	1.0	normalized
F_0	Nominal BF_3 flow	3.5	sccm
f_{11}	^{11}B fraction	0.80	—
$V_{x,0}$	Extraction voltage	10	kV
$V_{s,0}$	Suppression magnitude	2.0	kV
$I_{E,0}$	Extraction current, nominal	35	mA
σ_m	Resolving slit window	0.35	u
η_{tr}	Analyzer transmission	0.92	—
I_{par}	Parasitic extraction current	1.8	mA
$T_{f,0}$	Filament temperature	2600	K
$T_{c,0}$	Cathode temperature	2200	K
T_w	Wall temperature (hidden)	500	K
$T_{c,max}$	Overtemperature trip	2560	K
β_R	Richardson exponent	9.0	—
τ_f	Filament thermal	4.0	s
τ_c	Cathode thermal	1.5	s
τ_g	Gas response	0.08	s
τ_e	Electron response	0.002	s
τ_i	Ion species response	0.010	s
τ_{arc}	Discharge response	0.005	s
τ_{Ib}	Bias supply (hidden)	0.005	s
τ_{If}	Filament supply (hidden)	0.05	s
$T_{e,0}$	Electron temperature	6.0	eV
τ_{HV}	Normal supply response	1.0	ms
τ_{coll}	Collapse time constant	0.8	ms
R_f	Fault path resistance	5.0	kΩ
I_{max}	Supply current limit	120	mA
λ_0	Baseline glitch hazard	0.004	1/s
N_{st}	Scan strips (fixed)	32	—
n_{pass}	Passes per wafer	6	—
t_{scan}	Time per pass	2.0	s
e_D^{max}	Mean dose tolerance	1.5	%
U_D^{max}	Strip non-uniformity limit	1.0	%
σ_{I_E}	Extraction measurement noise	0.12	mA
σ_{I_B}	Faraday noise	0.05	mA
τ_{meas}	Measurement filter (hidden)	2.0	ms

8.3.9 Assumptions

This is a lumped control model. It is not a particle-in-cell code, not a Boltzmann solver, not a plasma-chemistry model and not a beam transport simulation, and its purpose is the signs, the nonlinearities, the time-scale separation and the fault logic.

The five species use effective yields and appearance-energy thresholds. The thresholds are published values used to order the onsets, and the yields are chosen so that the nominal recipe produces a plausible $^{11}B^+$ beam. Absolute species fractions are not predicted, and multiply charged ions, neutral fragments, charge exchange and wall recombination are absent.

A glitch is a switched conductance across a supply with a current limit. Real arcs depend on geometry, contamination, pressure, surface state, supply topology, electromagnetic coupling and the

protection hardware itself.

Hardware protection in a real implanter acts locally and far faster than any supervisory controller. The application demonstrates the sequencing and its control consequences, and every threshold, dwell and retry count needs equipment-maker validation before it means anything.

Signal names vary between equipment makers. The cathode shield in Figure 8.1 is passive and protective, and the electrically active electrode opposite the cathode is labeled repeller, which is one of several conventions in use.

The dose map is one dimensional with a single beam-enabled indicator. Two dimensional scan patterns, beam profile, neutralization, wafer charging, angle control and channeling are outside the model, and the twelve second wafer is a compression of a real implant so that faults and wafers happen on comparable time scales in a classroom.

8.4 Controllers as Implemented

Three loops on three clocks, and a supervisor that owns the permissives.

8.4.1 The Inner Loop and Its Two Gates

The inner loop regulates extraction current with cathode bias current on a scan of $\Delta t_c = 1\,\text{ms}$. It is written in the incremental, or velocity, form with a feedforward bias $I_{ff} = I_{b,0}$, and the complete law as coded is Equations (8.15) to (8.20).

$$e_k = I_{E,sp,k} - \hat{I}_{E,k} \tag{8.15}$$

$$s_k = s_{k-1} + K_p\,(e_k - e_{k-1}) \tag{8.16}$$

$$c_k = I_{ff} + s_k + z_{k-1}, \qquad u_k = \text{clip}\left(c_k, 0, I_{b,max}\right) \tag{8.17}$$

$$I_{b,cmd,k} = q_{auto}\,u_k + (1 - q_{auto})\,I_{b,hold} \tag{8.18}$$

Here $\hat{I}_E$ is the extraction-current reading after measurement noise and the first order filter τ_{meas}. Away from the gates this is the velocity form of Section 1.6.2 with no derivative term. The running sum s_k is S_k of Equation (1.13), the demand c_k is v_k of Equation (1.14) with the bias $u_b = I_{ff}$, and u_k is that demand clipped to $I_{b,max} = 4.0\,\text{A}$. Extraction current rises with bias current, so the loop is reverse acting in the convention of that section. While q_{auto} is one, $I_{b,hold}$ follows $I_{b,cmd}$. Equation (8.18) is a switch and not a blend. While the permissive is zero the command *is* the held value, so neither the proportional term nor the integrator can reach the actuator. Blending them with a gain instead would let a large proportional term leak through at exactly the moment the error is meaningless.

The integral state is updated after the output, and it carries three regimes in one recursion, Equation (8.19), after which it is clipped to $\pm 3 I_{b,max}$.

$$\begin{aligned} z_k = z_{k-1} + \Delta t_c \big[& q_{auto}\,(K_i e_k + K_{aw}(u_k - c_k)) \\ & + q_{track} K_t \left(z_{track,k} - z_{k-1}\right) \big] \end{aligned} \tag{8.19}$$

$$z_{track,k} = I_{b,hold} - I_{ff} - s_k \tag{8.20}$$

Running, $q_{auto} = 1$ and $q_{track} = 0$ give ordinary integral action with back-calculation for saturation, Equation (1.15) with $z = I$, $K_i = K_c/\tau_I$ and $K_{aw} = 1/T_t$. The default $K_{aw} = 5/\text{s}$ is a tracking time of 0.2 s, so each scan removes half a percent of the clipped excess. Held, both permissives are zero and the integrator is frozen, which back-calculation alone would not do. Recovering, the tracking permissive drives z with $K_t = 40/\text{s}$ toward the value that reproduces the held output at the present error, so the first automatic output equals the held output to within the measurement noise, 2.4944 A against 2.4930 A in scenario 7, and the transfer back is bumpless by construction rather than by luck [6].

8.4.2 The Outer Trim and the Filament Supervisor

The selected-beam loop trims the extraction setpoint on a 20 ms scan, Equation (8.21).

$$\dot{I}_{E,trim} = q_B \,\mathrm{clip}\left[K_{i,B}\left(I_{B,sp} - I_{B,meas}\right), -r_B, r_B\right] \tag{8.21}$$

with $K_{i,B} = 0.08/\mathrm{s}$, $r_B = 2.0\,\mathrm{mA/s}$, the trim bounded to $\pm I_{trim,max} = \pm 8\,\mathrm{mA}$ and the setpoint $I_{E,sp} = \mathrm{clip}(I_{E,base} + I_{E,trim})$ within 10–80 mA. The permissive q_B requires a valid beam measurement, the NORMAL state, and a ratio of selected to total current between 0.02 and 0.90, which is the cheapest available diagnostic for a beamline fault masquerading as a source fault. Selected beam rises with the extraction setpoint, so the trim is reverse acting like the inner loop.

The filament supervisor runs every 100 ms, acts only in NORMAL and only on a persistent bias offset, $\dot{I}_{f,sp} = K_f(I_{b,cmd} - I_{b,mid})$ with $I_{b,mid} = 2.25\,\mathrm{A}$, is rate limited to 0.5 A per execution and bounded to 100–190 A, and is frozen under the naive architecture and whenever the cathode is within 60 K of its trip. At the shipped gain $K_f = 0.08/\mathrm{s}$ and a bias command of 2.5 A it moves two hundredths of an ampere per second, which is two orders of magnitude slower than the inner loop. Bias current and filament current are two actuators on one variable and the only thing keeping them apart is that factor. A hotter filament lowers the bias command the inner loop needs, so the supervisor, which raises the filament current when the bias command sits above its center, is direct acting in the convention of Section 1.6.2.

🕮 **Pause.** Try Exercise 8.1 before reading on.

8.4.3 The Detector and the Supervisor

Detection requires coincidence on either electrode, Equation (8.22).

$$g = \bigvee_{k\in\{x,s\}} \Big(\left[I_{HV} > 0.6 I_{max}\right] \wedge \left[U_k < 0.75 U_{k,sp}\right] \wedge \left[t_{p,k} > t_{detect}\right] \Big) \tag{8.22}$$

Low beam current on its own is never sufficient and is never labeled an arc. The supervisor is a state machine over OFF, WARMUP, NORMAL, QUENCH, RECOVER, QUALIFY, REPAINT and TRIP. On detection it stores the last good command, records the scan coordinate, inhibits high voltage and enters QUENCH for the dwell t_q. It then ramps the affected supplies over t_r, qualifies the recovered extraction current inside eight percent of setpoint for a dwell t_{qual}, and resumes. If the committed dose deficit exceeds the repaint threshold it enters REPAINT first. More than n_{storm} events in a rolling ten seconds, or a cathode overtemperature, trips the source and the restart is an operator action rather than a timer.

Three architectures are selectable and they differ only in the permissives. A naive proportional-integral loop keeps q_{auto} at one throughout and has no outer trim. A cascade adds the trim and the filament supervisor, the structure Volume 3, Chapter 2 matches to where a disturbance enters, and still keeps q_{auto} at one. The arc-aware cascade gates both, and it alone repaints.

8.5 Two Judgments That Are Usually One

The application keeps two judgments apart that are normally made together, and the separation is the most transferable idea in it.

Latching a fault quenches the beam, so a false latch is expensive and the rule that triggers it demands corroborating evidence and a persistence time. Deciding that one measurement has stopped meaning anything is a weaker and much faster question. An extraction-current reading taken while the extraction voltage is not present is meaningless on its own terms, and waiting the coincidence window before saying so costs the controller a scan.

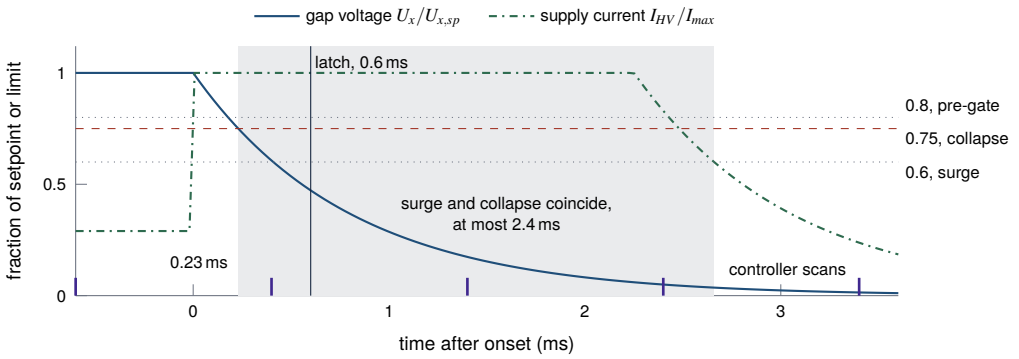


Figure 8.2 A spark against the controller's clock. The curves follow Equations (8.12) and (8.13) with the parameters of Table 8.2, the beam load left out after onset; only the printed times are the chapter's. The gap voltage (solid) passes the dashed 75 % line of the detector 0.23 ms after onset, and the fault latches at 0.6 ms. Notice the scan that falls between the two: it reads a falling extraction current before any latch, which is why the 80 % pre-gate (upper dotted line) needs no persistence. The supply current (dash-dotted) stays on its limit only while the gap still holds voltage, so the shaded coincidence lasts at most 2.4 ms.

That scan is not free. The gap voltage falls through seventy-five percent of its setpoint 0.23 ms after onset, which is the second integration step, and the coincidence rule needs 0.3 ms more, so the fault is latched 0.6 ms after onset. The 1 ms controller scan keeps its own clock. When a spark starts 0.2 or 0.4 ms before a scan, that scan reads a filtered extraction current that is already falling while nothing has been latched, and holding the command in force at the latch would freeze whatever that scan asked for. The application therefore keeps two defenses. A fast pre-gate marks the extraction-current reading invalid whenever the extraction voltage is below eighty percent of its setpoint, which needs no persistence because it latches nothing, and the command is snapshotted at the integration step whenever the electrical evidence is healthy, so that the snapshot, not the command at detection, becomes $I_{b,hold}$. Figure 8.2 draws that sequence against the controller's scans.

The consequence is measurable on a counterfactual copy of the controller with both defenses removed, firing the same fifteen millisecond spark at each of the five integration-step phases of the scan. With the onset 0.4 ms before a scan, the held command is 2.740 A against a pre-fault 2.492 A, the cathode gains 79 K over the following second, and the extraction current climbs to 57 mA, so the 150 ms qualification dwell is never completed and the supervisor is still in QUALIFY nine seconds later. At the other four phases the hold is within 0.018 A of the pre-fault value. Either defense alone, and the application as shipped, hold the command within 0.002 A of its pre-fault value at all five phases, and the cathode moves 0.15 K. The scripted spark of scenario 7 starts 0.2 ms after a scan and is latched before the next one, so on its own it would not expose the difference.

The same separation explains why a dead Faraday cup and a spark are handled differently. Two measurements carry two validity flags. The extraction-current flag is gated on high-voltage health, the selected-beam flag is gated on its own instrument, and a single global fault bit would either freeze the inner loop for no reason or walk the outer one on a dead reading.

8.6 Results

Every number below comes from the application's own engine at the default seed, 20260921 (Section 1.6.1), with the scenario named beside it. Scorecard figures are quoted as the scorecard computes them, and peaks, crossings and held values are read at the 0.2 ms integration step. Table 8.3 lists the measured outcomes. Where noise could matter, the case was rerun with other seeds and their range is given beside the shipped result.

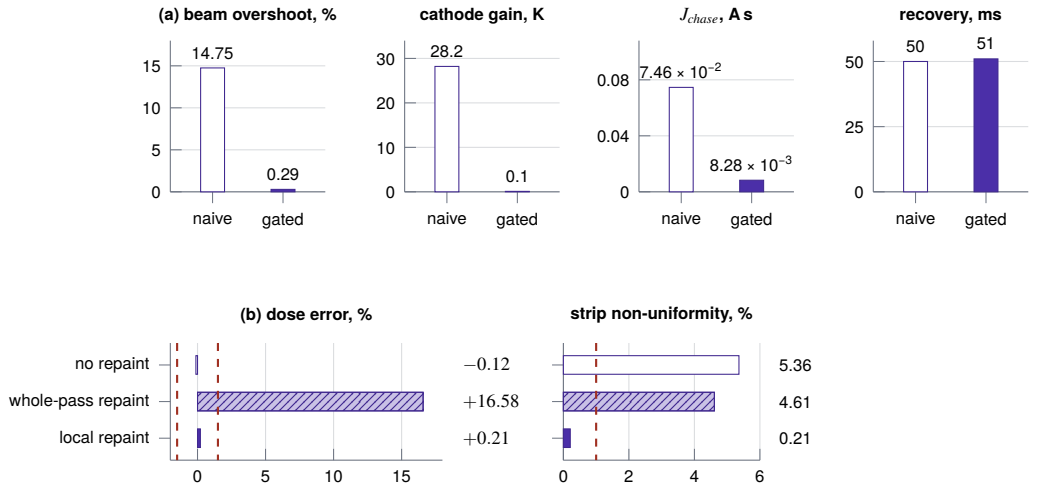


Figure 8.3 Two rows of Table 8.3 as charts. (a) Scenarios 6 and 7, the same fifteen millisecond spark on the naive PI (open bars) and on the gated, arc-aware cascade (solid bars). (b) Scenario 10, no repaint (open), a whole-pass repaint (hatched) and a local repaint (solid), with the mean dose tolerance of 1.5 % and the strip non-uniformity limit of 1.0 % of Table 8.2 dashed. Notice the last panel of (a): the gate removes the overshoot, the heat and the chasing, and leaves recovery where it was.

The twelve scenarios fall into two halves and a challenge. Scenarios 1 through 5 build the source and the cascade, from a cold start through the emission knee, the four currents, the mass spectrum and the outer trim to tuning without faults. Scenarios 6 through 11 are the fault, organized as contrasts rather than as a list of features. Scenario 6 against 7 is the gate, scenario 8 against 9 is specificity, scenario 10 is the product and scenario 11 is the retry limit. Scenario 12 is the graded challenge. Each scenario carries a completion rule in the app, and the Exercises at the end of the chapter are built on several of them. Figure 8.3 draws the two central contrasts of the table, one spark with and without the gate and the three repaint policies.

8.6.1 The Plant Before the Fault

The nominal recipe is a fixed point of every state, and the three currents an operator reads are 10.0 A of source discharge, 35.00 mA of extraction and 7.918 mA of mass-selected beam. The plasma supplies 33.20 mA against a space-charge limit of 55.45 mA, so the extraction voltage is not a current knob.

The nonlinearity that decides the architecture is the emission knee. Between 2090 K and 2420 K the normalized emission runs from 0.562 to 2.742, a factor of 4.9, and its slope with temperature changes by a factor of 3.7 across those three hundred and thirty kelvin, so a loop closed on a thermal variable has no single gain. Closing on extraction current instead puts the nonlinearity inside the loop where the integrator absorbs it.

Scenario 4 is the pair of disturbances that one current measurement cannot separate. Dropping the analyzer transmission by twelve percent moves the selected beam by −12.00 % and the total extraction current not at all. One second later the outer trim has raised the extraction current by 0.25 % and the beam is still 11.76 % low. Shifting the plasma toward fewer B^+ at 22 s moves the beam by a further −9.15 % without moving the total at all. An outer loop with unlimited authority reading only the Faraday cup would drive the source harder in both cases, which is why the trim is rate limited, bounded to ±8 mA and gated on the current ratio. At the shipped trim gain it is also slow. By 22 s it has recovered 2.7 of the twelve points the detune took and has used 1.04 mA of its authority, so at that gain the bound is never reached. The case's rerun raises the trim gain to 0.45/s.

Table 8.3 Measured outcomes at the default seed

Case	Configuration	Metric	Value
2	nominal recipe	I_{arc}, I_E, I_B	10.0 A, 35.00, 7.918 mA
2	—	I_{pl}, I_{CL}	33.20, 55.45 mA
1	T_c 2090 to 2420 K	emission ratio	0.562 to 2.742
3	V_a 50 to 75 V	T_e	6.00 to 6.88 eV
3	—	B^+ fraction	32.3 to 39.3 %
4	η_{tr} 1.00 to 0.88	I_B, I_E at once	−12.00, 0.00 %
4	—	same, 1 s later	−11.76, +0.25 %
6	naive PI, 15 ms	$I_{b,cmd}$	2.493 to 4.0 A
6	—	time at limit	32 ms
6	—	ΔT_c, overshoot	28.2 K, 14.75 %
6	—	J_{chase}, recovery	7.46×10^{-2} A s, 50 ms
7	arc-aware, same spark	$I_{b,cmd}$ held	2.493 A
7	—	ΔT_c, overshoot	0.1 K, 0.29 %
7	—	J_{chase}, recovery	8.28×10^{-3} A s, 51 ms
8	flow to 60 %	I_B dip, $I_{b,cmd}$	6.62 mA, 3.378 A
8	—	fault state entered	none
9	60 ms dropouts	$I_{b,cmd}$, trim	2.489–2.511 A, 0 mA
10	no repaint	dose error, U_D	−0.12, 5.36 %
10	whole-pass repaint	dose error, U_D	+16.58, 4.61 %
10	local repaint	dose error, U_D	+0.21, 0.21 %
11	six events in 4.5 s	trip at, energy	3.6 s, 24.3 J
11	—	energy, settled	32.3 J
11	limit 20	energy, no trip	32.6 J
12	as given	worst I_B error	0.946 mA
12	suggested	worst I_B error	0.177 mA
12	as given, suggested	wafers in spec	0 of 5, 2 of 4

Half of the detune loss is then back 7.6 s after the detune, and after the species shift the trim reaches its 8 mA bound at 33.9 s, leaving the beam 0.50 mA short at 40 s.

8.6.2 Saturation Without Windup

Scenario 6 fires a fifteen millisecond extraction spark on a loop with textbook back-calculation anti-windup, the remedy Volume 2, Chapter 3 applies at a true actuator limit, and no gating. The bias command runs from 2.493 A to its 4.0 A limit, the cathode gains 28.2 K, the selected beam overshoots to 9.065 mA against a 7.9 mA setpoint, and the fault-chasing index $J = \int \max(I_b - I_{b,pre}, 0)\, dt$ over the two seconds after detection is 7.46×10^{-2} A s.

> **TRY IT IN THE APP** **6 · A Naïve PI Chases a Spark**
>
> Load the scenario and press Run. At 3 s a fifteen millisecond spark crosses the extraction gap. Watch the bias-current command climb from 2.493 A to its 4.0 A limit while the extraction voltage is on the floor, and the selected beam overshoot to 9.065 mA when the high voltage returns. Nothing about the source was wrong.

The word windup is slightly wrong for what happens. The proportional term alone asks for 2.80 A above the feedforward once the reading has fallen, so the output sits on its limit for 32 ms, from 3 ms after onset until the high voltage is ramping back. Back-calculation engages for exactly that interval and does what it was designed to do, holding the integral below 0.23 A. The damage is heat. Nearly 4 A of bias for those milliseconds raises the cathode 28.2 K over the next 60 ms, and it is that stored heat, not the integral, that overshoots the beam when the gap returns. The loop has spent its actuator on an error that does not exist.

Pause. Try Exercise 8.2 before reading on.

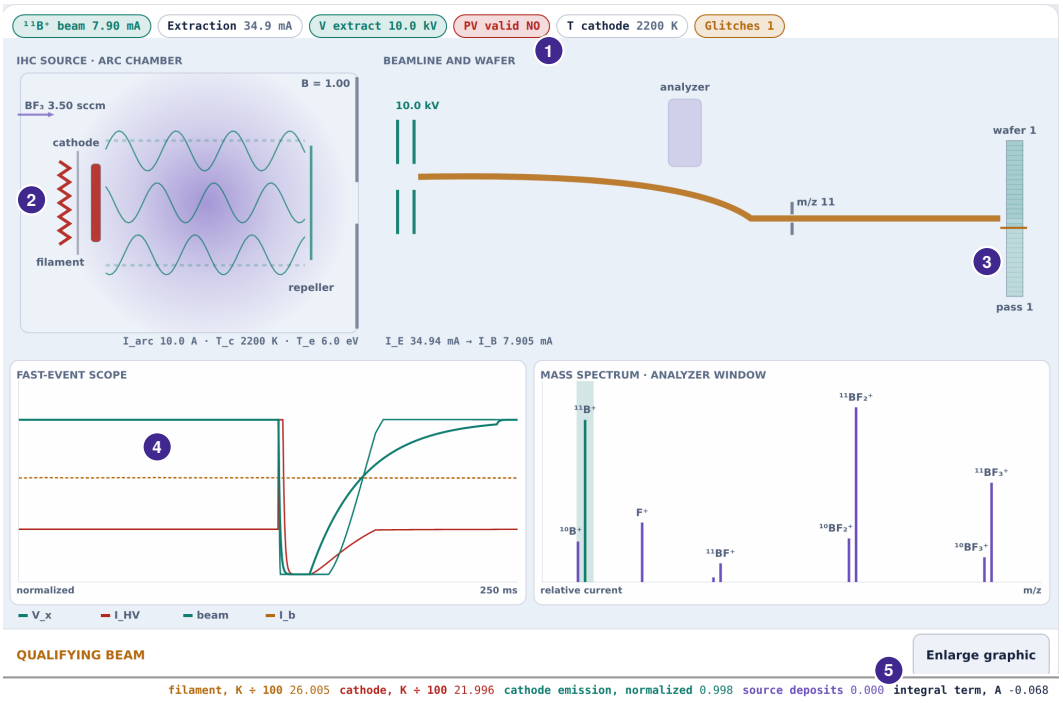


Figure 8.4 The application during scenario 7 at $t = 3.12$ s, 120 ms after a fifteen millisecond extraction spark under the arc-aware architecture, with the supervisor in QUALIFY (the banner under the plant view) and the extraction-current reading still marked invalid (1). The crop keeps the plant view and, under the gray rule, the labels of the thermal strip. The source cutaway (2) carries the filament, the cathode, the repeller and the electron helices whose brightness follows the electron density, and the beamline carries the bent beam into the resolving slit and the scanned wafer (3) with its strips colored by delivered dose. The fast-event display (4) below them holds the last 250 ms at the integration step, which is where the supply current reaching its limit and the gap voltage collapsing are legible while the bias current stays flat at its held 2.493 A. The extraction current reads 34.94 mA, the selected beam 7.905 mA and the fault-energy ledger 8.0 J. The integral term (5) is plotted on the thermal strip on purpose, because that is where a millisecond fault would otherwise be written into a four second plant. Here it reads −0.068 A as the tracking permissive rebuilds it before the loop is handed back.

Scenario 7 is the same gains, the same spark and the same seed with the measurement gated. The bias command is held at 2.493 A from the latch until the beam requalifies, the cathode gains 0.1 K, the overshoot falls to 0.29 % and J falls to 8.28×10^{-3} A s. Over seeds 1 to 5 the gated J runs from 1.3×10^{-3} to 4.8×10^{-3} A s against 6.9×10^{-2} to 7.3×10^{-2} A s ungated, because what remains with the gate is measurement noise rather than the spark. Recovery time is essentially unchanged, 51 ms against 50 ms, because recovery belongs to the high-voltage ramp and not to the controller. That last comparison is the one worth carrying away. Gating the measurement does not make the tool recover faster. It stops the tool from doing damage while it waits.

8.6.3 Specificity Is the Harder Half

A detector is easy to make sensitive. Scenarios 8 and 9 test the other direction. Dropping the real gas flow to sixty percent takes the beam from 7.92 mA to 6.62 mA, the inner loop raises the bias command to 3.378 A to restore the extraction current, and the supervisor never leaves NORMAL.

That is correct behavior, and a fault gate written to include low beam current would have frozen the controller through exactly the disturbance feedback exists to reject. The beam also settles above where it started, at 8.35 mA after thirty seconds, because less gas raises the electron temperature to 6.36 eV and enriches the B^+ fraction, so the same extraction current carries more boron.

Pause. Try Exercise 8.3 before reading on.

Killing the Faraday reading for 60 ms with every electrical quantity normal leaves the bias command inside the band its measurement noise gives it anywhere in the run, 2.489 to 2.511 A during the first dropout against 2.482 to 2.519 A over the whole ten seconds. The outer trim does not move at all, and again there is no quench. The inner loop keeps regulating because its own measurement never became invalid, and the outer trim holds its accumulated value because its measurement is the dead one.

DID YOU KNOW? Same Move, Opposite Verdicts

From the simulator: in Table 8.3 a real gas loss to 60 % drives the bias command to 3.378 A, and that is feedback doing its job. A fifteen millisecond spark drives it to its 4.0 A limit, and that is the fault. The size of the move does not tell them apart. Only the high-voltage evidence does.

8.6.4 Putting the Dose Back Where It Went Missing

An implant specification is an integral, so an interruption is a spatial defect. Scenario 10 fires a fifteen millisecond spark at 4 s, at the start of the third of six passes, and grades the finished wafer three ways. The beam is physically off for 26 ms, which leaves strip 0 at 0.897 of its committed dose. No repaint gives a dose error of −0.12 % and a strip non-uniformity of 5.36 %. A whole-pass repaint gives +16.58 % and 4.61 %. A local repaint gives +0.21 % and 0.21 %.

The whole-pass result is the instructive one. The extra pass starts where the scanner stands when qualification ends, at strip 4, and runs one full pass of travel, so every strip gains a sixth of its dose. Strip 0 ends at 1.064 against about 1.17 for its neighbors, exactly as far below them as before, and the mean ends 16.58 % high. The non-uniformity falls slightly only because it is normalized by a larger mean. Local repaint integrates the deficit of Equation (8.14) strip by strip with the beam blanked over strips already at their committed dose. It needs no record of where the scanner was at detection, because both ledgers are kept per strip and the deficit carries its own location.

8.6.5 Knowing When to Stop

Scenario 11 fires six events in 4.5 s, four on the extraction gap and two on the suppression gap, on a source running dirty. Each one is handled correctly, which is the problem. The tool trips 3.6 s into the storm on the fifth event with 24.3 J on the fault-energy ledger, and the restart is an operator action. A trip inhibits the high voltage but cannot recall a spark already conducting, so the ledger settles at 32.3 J. With the limit raised to 20 the tool restarts through all six and reads 32.6 J, because the only spark after the trip is a 0.32 J suppression spark against 8.0 J for each extraction spark. A trip saves only the sparks that come after it. Raising the limit is not merely permissive, because every restart puts the full extraction voltage back across a gap that has already broken down several times, and the fault-energy column is the honest measure of what that costs. The storm is scripted, so these figures do not change with the seed.

TRY IT IN THE APP **11 · An Arc Storm and a Safe Trip**

Press Run and watch the rolling event count against the Arc-storm trip limit. The source, running dirty, takes six sparks in 4.5 s, and each one is handled correctly. The tool still trips on the fifth, 3.6 s into the storm, and stays tripped. Restarting is an operator action: press Clear the trip and restart.

8.6.6 The Graded Run

The challenge is seventy seconds containing a gas calibration step, an analyzer detune at 12 s, three sparks of two lengths on two electrodes, an arc storm that trips the tool, an operator restart at 34 s and a source cleaning at 35 s, beginning from a single proportional-integral loop with a three millisecond detection window, a nearly dead outer trim at 0.02/s and no repaint. The model parameters and disturbances must stay as given. The run is scored on the worst selected-beam error after 46.2 s, the last third of the run, against a band of ± 0.35 mA around the 7.9 mA setpoint. As given, that error is 0.946 mA, 0.596 mA outside the band, and none of five wafers meets specification. The suggested settings correct the architecture to arc-aware, the detection window to 0.3 ms, the trim gain to 0.45/s and the repaint policy to local, and retune the inner loop to $K_p = 0.10$ A/mA and $K_i = 0.45$ A/(mA s). With them the worst error is 0.177 mA, inside the band with 0.173 mA to spare, and two of four wafers pass. Seeds 1 to 5 give 0.938 to 0.967 mA as given and 0.164 to 0.188 mA with the suggested settings.

The detection window does not do what its three millisecond value suggests. The coincidence of surge and collapse lasts at most 2.4 ms for any spark, because the supply current falls below sixty percent of its limit once the gap voltage has collapsed, so at 3 ms none of the nine sparks in the run is latched, not merely the short ones. Under the arc-aware architecture the fast pre-gate then holds the output only while the extraction voltage is below eighty percent of its setpoint, and the loop chases the tail of the filtered reading to 3.29 A of bias at the 8 s spark, against the limit for the naive loop. What decides the score is the trim gain, because the error left in the scored window is the 12 s detune, which only the selected-beam loop can see. The cascade with the trim gain at 0.45/s and nothing else changed already scores 0.117 mA, while the arc-aware loop with the 0.3 ms window and the shipped trim scores 0.828 mA. The gate, the window and the repaint are visible in the fault-chasing index and the dose ledger instead. Shortening the window alone puts the naive loop's J at 0.31 A s over the run, and the suggested settings bring it to 0.012 A s. Shortening it costs nothing in nuisance trips, because the rule is a coincidence of surge and collapse rather than a threshold on any one signal.

8.7 Technology in Practice

The three loops of Section 8.4 read the current of the extraction supply and a Faraday cup at the end of the beamline, and they move the bias and filament supplies of the source. The detector reads the same supply current beside the gap voltage, and the sequence acts on the extraction supply itself. Figure 8.5 draws them with their tags.

What Measures It

The selected beam, the quantity the wafer is paid for, is read on a Faraday cup.

TECHNOLOGY **Sensor: A Faraday Cup**

A Faraday cup is an insulated metal cup that stops the beam, wired to a current-sensitive amplifier, so it measures the charge the beam carries directly. Every ion that strikes it frees secondary electrons with energies below about 10 eV, and any that escape make the reading wrong by their current. A long cup, a suppression voltage at the entrance or a permanent magnet keeps them in. Because it stops the beam, the cup must also take the beam's energy as heat [19]. In an implanter a cup samples the scanned beam during the implant, a current integrator sums the charge into dose, and the cup is calibrated before the first wafer of a batch. The scan must overscan the cup completely, or the dose varies [20].

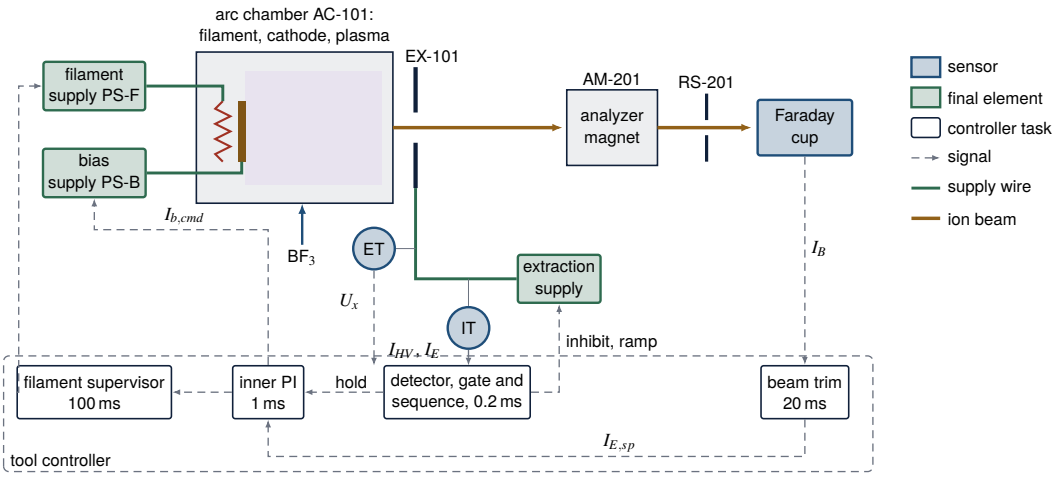


Figure 8.5 The cascade of Figure 8.1 drawn with its hardware. The current transmitter IT on the extraction supply serves twice: the detector reads it as supply current beside the gap voltage ET, and the inner loop receives it as extraction current only through the gate. The Faraday cup, which samples the scanned beam beside the wafer, feeds the beam trim alone. Notice that the detector moves no source supply: it gates a measurement, holds the inner loop and inhibits or ramps the extraction supply.

The model carries the cup as a reading with 0.05 mA of noise (Table 8.2). In scenario 9 it goes to zero for 60 ms while every electrical quantity stays normal, which is how a failed cup or its amplifier looks on a trend, and why the trim has a validity flag of its own.

What Moves It

The inner loop commands the bias supply between 0 and 4.0 A and the supervisor the filament supply between 100 and 190 A.

> **TECHNOLOGY** **Final Element: The Supplies Around the Cathode**
>
> An indirectly heated cathode is driven by three supplies. One sends current through the filament, one biases the filament against the cathode cap so that its electrons bombard and heat the cap, and an arc supply holds the cathode 50 to 150 V negative of the chamber walls. More bias means more energetic electrons striking the cap and more arc current. A design patented in 1996 closed two loops on these supplies. The arc current, compared with a demand the implant controller set from the desired beam current, adjusted the bias supply, and the cathode's emission current adjusted the filament supply. Shielding the filament behind the cap kept it out of the plasma and extended its life [21].

The model keeps the same pair of supplies as first-order lags of 5 ms and 50 ms, with the bias capped by what the filament can emit. Its fast loop closes on extraction current rather than arc current, and its filament supervisor plays the slow loop.

Where the Controller Runs

The loops belong to the implanter's own control system. In a 1996 tool a workstation was linked to a bus controller whose device interfaces drove the sensors and actuators, and recipes set the species, energy, beam current and dose for an automatic set-up [20]. Control loops run with cycle times from milliseconds to minutes [22]. The app runs the detector, the sequence and the dose ledger every 0.2 ms, the inner loop every 1 ms, the trim every 20 ms and the supervisor every 100 ms. Real arc

protection acts in the supply hardware, faster still. The structure of this chapter is a velocity-form PI block, two slow trims and a state machine that owns their permissives.

> **IN PRACTICE** **Counting Glitches Before a Tool Is Accepted**
>
> In a 1996 qualification of a medium-current implanter, the unscanned beam ran for two hours on a chart recorder. A glitch, a sudden drop-out or spike of more than ten percent of the beam current caused by an arc or another upset in the source, could occur at most ten times in those two hours, and fewer than five times an hour in the later qualification tests. Dose uniformity and repeatability were read on monitor wafers by four-point probe. Repeatability was the hardest to reach: filament current, arc current and extraction settings drift as the source ages [20].

What Is Modernizing It

Three developments bear on this source. Deposits cause premature failures of the source, and in-situ chemical cleaning cycles improved beam stability, glitch rate and source lifetime on a production tool compared with a tool without them [23]. Predictive maintenance trained on production data forecasts the breaking of the filament [24]. And because an implanted wafer cannot be reworked, classifiers watch tool data drawn through the equipment's SECS-II interface to detect faults in real time [25]. The deposit state of this chapter, which raises the glitch hazard, is the variable the first of these acts on.

8.8 Limitations

The plasma model is lumped and reduced. It reproduces the ordering of the species onsets and the bounded effect of confinement, and it does not predict absolute species fractions, plasma potential, sheath structure or contamination.

The glitch is a teaching model of a switched conductance. Real protection hardware acts locally and faster than anything in this application, and every threshold, dwell, retry count and ramp order requires validation by the equipment maker before it is used for anything.

The deposit state carries a hazard rate that rises with deposits and extraction voltage. It exists to make faults more likely in degraded scenarios and must never be read as a predicted time to arc.

The historian logs at 100 Hz, so the strip charts carry one or two samples of a fifteen millisecond event and none of a two millisecond one. The sub-millisecond shape is available only on the fast-event display in the scene, and the export does not carry high-rate rows around a fault.

The mass analyzer is an empirical mass window. Beam optics, emittance, neutralization, wafer charging, angle control, channeling, junction depth, activation and crystal damage are all outside the scope, and repaint restores a ledger rather than proving device-level equivalence after a real interruption.

> **OPEN CHALLENGE** **A Source That Lasts on One Dopant**
>
> Replacing the ion source is generally the most frequent maintenance on an implanter [26]. Fabs now expect stable beams with minimal arcing, at least 400 hours of source life and more than 90 % uptime [27]. Life is getting harder to reach. In advanced logic many doping steps have moved into epitaxy, so a tool runs one species for long stretches, which shortens source life. A co-gas with hardware changes doubled or tripled it for some species [11]. Progress would be a source that keeps its life and its arc rate on any single species. The deposit state of this chapter only raises a hazard rate and predicts no arc.

8.9 Conclusion

This chapter studied an indirectly heated cathode boron trifluoride ion source on an interactive simulator in which a three-level cascade regulates a beam whose accelerating gap sparks. Sixteen continuous states carry two thermal masses, two supply currents, the gas and electron densities, the source discharge, five ion species, two electrode voltages, a deposit state and a fault-energy ledger,

and the fault detection and the dose ledger run at the integration step because protection cannot wait for a controller scan.

The measured outcome that organizes the application is what one decision is worth. The same gains and the same fifteen millisecond spark drive the bias command to its 4.0 A limit and the beam 14.75 % past setpoint without a measurement-quality gate, and hold the command where it was and the beam within 0.29 % with one. Recovery time is unchanged at about 50 ms in both cases, because recovery belongs to the high-voltage ramp. The rule that separates them is short. Record the error, because the wafer really did miss dose, and do not feed it to the loop, because the source was never the reason.

KEY POINTS

- A spark collapses the beam reading for milliseconds while the source is fine, so a loop that believes it heats the source for nothing.
- Back-calculation does not prevent it: the proportional term alone drives the bias to its 4.0 A limit, and the cathode stores the heat.
- Gating the measurement, holding the output and freezing the integrator cut the overshoot from 14.75 to 0.29 %, with recovery unchanged near 50 ms.
- Latching a fault needs corroborating evidence: low beam current on its own is never sufficient.
- A local repaint driven by a per-strip ledger returns the wafer to 0.21 % non-uniformity, where a whole-pass repaint overdoses it by 16.58 %.

Where else this applies. A protective override that takes a loop outside its domain in Volume 4, Chapter 2, windup at a true actuator limit in Volume 2, Chapter 3, and a product record kept apart from the instruments in Volume 8, Chapter 13.

References

[1] R. B. Fair, "History of some early developments in ion-implantation technology leading to silicon transistor manufacturing," *Proceedings of the IEEE*, vol. 86, no. 1, pp. 111–137, 1998.

[2] C. B. Yarling, "History of industrial and commercial ion implantation 1906–1978," *Journal of Vacuum Science & Technology A*, vol. 18, no. 4, pp. 1746–1750, 2000.

[3] B. A. MacKinnon and J. P. Ruffell, "Semiconductor ion implanters," in *Application of Accelerators in Research and Industry: Twenty-First International Conference*, ser. AIP Conference Proceedings. AIP, 2011, pp. 46–51.

[4] K. J. Åström and L. Rundqwist, "Integrator windup and how to avoid it," in *1989 American Control Conference*. IEEE, 1989, pp. 1693–1698.

[5] M. V. Kothare, P. J. Campo, M. Morari, and C. N. Nett, "A unified framework for the study of anti-windup designs," *Automatica*, vol. 30, no. 12, pp. 1869–1883, 1994.

[6] L. Zaccarian and A. R. Teel, "A common framework for anti-windup, bumpless transfer and reliable designs," *Automatica*, vol. 38, no. 10, pp. 1735–1744, 2002.

[7] M. Blanke, M. Kinnaert, J. Lunze, and M. Staroswiecki, *Diagnosis and Fault-Tolerant Control*, 3rd ed. Berlin, Germany: Springer, 2016.

[8] J. H. Freeman, "A new ion source for electromagnetic isotope separators," *Nuclear Instruments and Methods*, vol. 22, pp. 306–316, 1963.

[9] T. N. Horsky, "Indirectly heated cathode arc discharge source for ion implantation of semiconductors," *Review of Scientific Instruments*, vol. 69, no. 2, pp. 840–842, 1998.

[10] J. N. Mefo, B. J. Sealy, E. J. H. Collart, D. G. Armour, and R. M. Gwilliam, "Characterisation of an indirectly heated cathode source using langmuir probe diagnostics," *Nuclear Instruments and Methods in Physics Research Section B*, vol. 240, no. 1-2, pp. 13–17, 2005.

[11] E. Su, A. Chang, K. Wu, and J. Tsai, "Extension of the source lifetime in HC ion implanter with dedicated species," in *2016 21st International Conference on Ion Implantation Technology (IIT)*. IEEE, 2016.

[12] I. N. Draganic, C. P. Leibman, L. J. Rybarcyk, T. M. Bronson, and K. W. Dudeck, "Hot filament performance in a Freeman ion source," in *Proceedings of the 17th International Conference on Ion Sources*, ser. AIP Conference Proceedings, vol. 2011. AIP Publishing, 2018, p. 030005.

[13] D. J. Chivers, "Freeman ion source: An overview (invited)," *Review of Scientific Instruments*, vol. 63, no. 4, pp. 2501–2506, 1992.

[14] C. D. Child, "Discharge from hot CaO," *Physical Review (Series I)*, vol. 32, no. 5, pp. 492–511, 1911.

[15] I. Langmuir, "The effect of space charge and residual gases on thermionic currents in high vacuum," *Physical Review*, vol. 2, no. 6, pp. 450–486, 1913.

[16] S. Dushman, "Electron emission from metals as a function of temperature," *Physical Review*, vol. 21, no. 6, pp. 623–636, 1923.

[17] J. C. Olson, A. Renau, and J. Buff, "Scanned beam uniformity control in the VIISta 810 ion implanter," in *1998 International Conference on Ion Implantation Technology Proceedings*, vol. 1. IEEE, 1999, pp. 169–172.

[18] T. F. Edgar, S. W. Butler, W. J. Campbell, C. Pfeiffer, C. Bode, S. B. Hwang, K. S. Balakrishnan, and J. Hahn, "Automatic control in microelectronics manufacturing: Practices, challenges, and possibilities," *Automatica*, vol. 36, no. 11, pp. 1567–1603, 2000.

[19] P. Forck, "Lecture notes on beam instrumentation and diagnostics," Joint University Accelerator School and GSI, Darmstadt, Germany, Lecture notes, 2011. [Online]. Available: https://www-bd.gsi.de/conf/juas/juas_script.pdf

[20] S. K. Joung, "Qualification of a medium current ion implantation system in a semiconductor production environment," S.M. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1996. [Online]. Available: https://hdl.handle.net/1721.1/40220

[21] P. Sferlazzo, E. K. McIntyre, Jr., W. E. Reynolds, R. M. Cloutier, and T. N. Horsky, "Ion generating source for use in an ion implanter," U.S. Patent 5,497,006, filed 15 November 1994, issued 5 March 1996, 1996. [Online]. Available: https://patents.google.com/patent/US5497006A/en

[22] K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, A. Hahn, S. Saravia, A. Sherule, and M. Thompson, "Guide to operational technology (OT) security," National Institute of Standards and Technology, Gaithersburg, MD, Special Publication 800-82r3, 2023.

[23] T. Romig, D. Newman, M. Mitchell, K. Ditzler, D. Eldridge, S. Bishop, and J. Mayer, "Implanter source life and stability improvement using in-situ chemical cleaning," in *Ion Implantation Technology: 17th International Conference on Ion Implantation Technology*, ser. AIP Conference Proceedings. AIP, 2008, pp. 364–367.

[24] G. A. Susto, A. Schirru, S. Pampuri, D. Pagano, S. McLoone, and A. Beghi, "A predictive maintenance system for integral type faults based on support vector machines: An application to ion implantation," in *2013 IEEE International Conference on Automation Science and Engineering (CASE)*. IEEE, 2013, pp. 195–200.

[25] S. C. Horng and Y. L. Lin, "Fuzzy tree classification system for fault diagnosis on ion implanter," *Journal of Physics: Conference Series*, vol. 2020, no. 1, p. 012017, 2021.

[26] A. Uvais, M. Jinguji, Y. Sato, T. Yotsumoto, and A. Botet, "Extending ion source life on high current ion implant tools with in-situ chemical cleaning," in *Ion Implantation Technology 2010: 18th International Conference on Ion Implantation Technology*, ser. AIP Conference Proceedings. AIP, 2011.

[27] T.-J. Hsieh, "Ion implantation ion sources in semiconductor fabrication: source life, beam stability, defects, yield improvements and contamination control," *Journal of Physics: Conference Series*, vol. 3237, no. 1, p. 012076, 2026.

8.10 Exercises

8.1 ⌕ Concept check · 1 min

Four quantities in the implanter are called current. Which pairing describes the inner loop of the cascade, the one that runs on the 1 ms scan?

(a) It reads the selected beam I_B and moves the extraction-current setpoint

(b) It reads the cathode temperature T_c and moves the filament current I_f

(c) It reads the extraction current I_E and moves the cathode-bias current I_b

(d) It reads the bias command $I_{b,cmd}$ and moves the filament current I_f

8.2 ⌕ Concept check · 2 min

In scenario 6 the naive loop runs back-calculation anti-windup with $K_{aw} = 5/\mathrm{s}$, yet the selected beam overshoots by 14.75 % when the high voltage returns. Why does back-calculation not prevent the overshoot?

(a) The naive architecture runs the inner loop with back-calculation switched off
(b) The tracking permissive pulls the integral toward a corrupted held output
(c) The integral is clipped at ± 12 A, far too wide to bound anything
(d) The proportional term alone drives the bias to its limit, and the cathode stores that heat

8.3 Concept check · 2 min

In scenario 8 the real BF_3 flow falls to 60 % and the selected beam dips from 7.92 to 6.62 mA. Why does the arc-aware supervisor stay in NORMAL with the PV-valid flag at one?

(a) The gas loss raises no supply-current surge and collapses no electrode voltage
(b) The 0.3 ms persistence window is too short to see a slow disturbance
(c) The outer trim restores the beam before the detector finishes its count
(d) Low beam current latches a fault only when it outlasts the quench dwell

8.4 Theory and calculation · 6 min

Use the normalized emission of Equation (8.3) with $\beta_R = 9$, $T_{c,0} = 2200$ K and no deposits ($w = 0$).

(a) Compute $\tilde{I}_e$ at 2090, 2310 and 2420 K to three decimals. *Checkpoint:* 0.562, 1.692 and 2.742.

(b) Differentiate $\tilde{I}_e$ with respect to T_c and find the ratio of the slope at 2420 K to the slope at 2090 K, to one decimal. *Checkpoint:* 3.7.

(c) In scenario 6 the cathode gains 28.2 K during the spark. Treat the selected beam as proportional to $\tilde{I}_e$, which holds while extraction is plasma limited, and find the beam rise this predicts, in percent to one decimal. Compare it with the 14.75 % overshoot the chapter measures. *Checkpoint:* 15.0 %, within about a quarter of a point of the measured overshoot.

8.5 Theory and calculation · 8 min

A spark switches the fault conductance across the extraction gap. Use Equations (8.12), (8.13) and (8.22) with Table 8.2: U_x starts at 10 kV and decays toward zero with $\tau_{coll} = 0.8$ ms while the spark conducts, $R_f = 5.0\,\mathrm{k\Omega}$ and $I_{max} = 120$ mA.

(a) Find the time after onset at which U_x falls below 75 % of its setpoint and the time at which it halves, in milliseconds to two decimals. *Checkpoint:* 0.23 ms and 0.55 ms.

(b) Once the gap has collapsed the extraction current is near zero. Find the gap voltage below which the supply current falls under $0.6\,I_{max}$, the time after onset at which that happens, to two decimals, and so how long surge and collapse coincide. Does a 3 ms persistence window latch a 15 ms spark? *Checkpoint:* 0.36 kV at 2.66 ms, so about 2.43 ms of coincidence. No.

(c) Integrate $\dot{E}_{fault}$ over a spark that lasts many τ_{coll} to find the energy of one extraction spark and of one suppression spark at $U_s = 2.0$ kV. Scenario 11 fires four extraction and two suppression sparks. Find the ledger with no trip and compare it with the chapter. *Checkpoint:* 8.0 J, 0.32 J and 32.64 J, against 32.6 J measured.

8.6 Theory and calculation · 7 min

In scenario 4 the analyzer transmission falls from 1.00 to 0.88 at 6 s. The inner loop holds I_E on its setpoint, and the selected beam is proportional to the ion part of the extraction current, $I_E - I_{par}$ at nominal extraction voltage (Equation (8.10)), with 7.918 mA of beam for 33.20 mA of ions before the detune.

(a) Find the incremental gain $g = \partial I_B / \partial I_{E,sp}$ after the detune, in milliamperes of beam per milliampere, to three decimals. *Checkpoint:* 0.210.

(b) Ignoring the rate limit r_B, Equation (8.21) then makes the beam error decay as a first-order lag. Find its time constant at $K_{i,B} = 0.080/\mathrm{s}$ and the fraction of the detune loss recovered 16 s after the detune, in percent to the nearest percent. *Checkpoint:* 59.6 s and 24 %.

(c) Find the time constant and the half-recovery time at $K_{i,B} = 0.45/\mathrm{s}$, to one decimal, and show that the rate limit stays inactive. Compare with the chapter. *Checkpoint:* 10.6 s and 7.3 s, against 7.6 s measured.

8.7 App challenge · 8 min

Load scenario 6, "A naïve PI chases a spark" (default seed), and run it at 1 s/s to 6 s. The scorecard tiles stop changing 2 s after the spark is latched, so any reading after 5.1 s is the same. Then load scenario 7, "The same spark, gated", and run it to 6 s.

(a) Read the J chase, Overshoot and Recovery tiles after each run. *Checkpoint:* 0.0746 A s, 14.8 % and 50 ms. Then 0.0083 A s, 0.3 % and 51 ms.

(b) In scenario 7 press Reset, set Quench dwell to 40 ms and Beam qualification dwell to 500 ms, and run to 6 s again. Read the Recovery tile. Export the CSV and find the first row whose `supervisor_state` reads `repaint`, and the same row in a run with the shipped dwells. *Checkpoint:* 77 ms. REPAINT begins at 3.64 s, against 3.26 s with the shipped dwells.

(c) The dwells grew by 380 ms in all, yet Recovery grew by only about 26 ms. Explain both numbers. *Checkpoint:* Recovery ends when the beam is first back within 2 % of setpoint, during RECOVER and QUALIFY. Only the quench dwell delays that.

8.8 App challenge · 8 min

Load scenario 10, "The dose hole, and what to do about it" (default seed), which fires a 15 ms spark at the start of the third of six passes with Dose repaint at Local. Run it at 1 s/s to 16 s, by which time the first wafer has finished, and export the CSV. Press Reset, set Dose repaint to None, run to 16 s and export again. Press Reset, set it to Whole pass, and repeat.

(a) In each file, find the first row with `wafers_done` equal to 1 and read its time, `metric_dose_error_pct` and `metric_uniformity_pct` to two decimals. *Checkpoint:* Local 13.78 s, +0.21 and 0.21 %. None 12.01 s, −0.12 and 5.36 %. Whole pass 14.01 s, +16.58 and 4.61 %.

(b) Non-uniformity is the strip range divided by twice the mean. Starting from the no-repaint wafer, predict the whole-pass dose error and non-uniformity if the extra pass adds one sixth of the committed dose to every strip, to one decimal. *Checkpoint:* +16.5 and 4.6 %.

(c) From the `supervisor_state` column, how long does each repaint last? Compare it with how much later each wafer finishes than the unrepaired one, and explain why the two agree to within one 10 ms row. *Checkpoint:* Local 1.78 s against 1.77 s, whole pass 2.00 s against 2.00 s. A repaint suspends the production pass instead of consuming it.

8.9 App challenge · 9 min

Load the challenge, "Challenge · a lot through an aging source" (default seed), and run it as given at 3 s/s to 70 s, where the checklist judges it.

(a) Read the checklist note and the J chase tile. *Checkpoint:* Worst 0.946 mA, 0 of 5 wafers in specification, 1 trip. J chase 0.0000 A s.

(b) Press Reset, set Control architecture to Cascade and Selected-beam trim gain to 0.450, and run to 70 s. The item now passes and the note is gone, so export the CSV. Its time column is in minutes. Over the rows from 0.77 to 1.1667 min (46.2 to 70 s), find the largest $|$`beam_meas_mA` − `beam_sp_mA`$|$, and read `wafers_good` and `wafers_done` in the row at 1.1667 min. *Checkpoint:* 0.117 mA, 2 of 5.

(c) Press Reset, set the architecture back to Naïve PI, the trim gain back to 0.020 and Coincidence persistence to 0.30 ms, and run to 70 s. Read the checklist note and the J chase tile, and explain why the tile read zero in part (a) and why the score did not move. *Checkpoint:* Worst 0.946 mA, 0 of 5. J chase 0.3139 A s.

8.10 Application area · 4 min

An implanter loses beam and the cause has to be named before anyone acts.

(a) For an extraction spark, a suppression spark, a loss of BF_3 flow and a dead Faraday cup, say which of these happen before any protection acts: a surge of the high-voltage supply current, a collapse of the extraction voltage, a collapse of the suppression voltage, and a fall of the selected-beam reading. *Checkpoint:* Extraction spark: surge, extraction collapse, beam reading. Suppression spark: surge, suppression collapse. Gas loss: beam reading only. Dead cup: beam reading only.

(b) Give the correct response to each, and say which of them is a protection event. *Checkpoint:* Both sparks: latch, quench, hold and freeze. Gas loss: let feedback act. Dead cup: freeze the trim, keep the inner loop.

8.11 Application area · 3 min

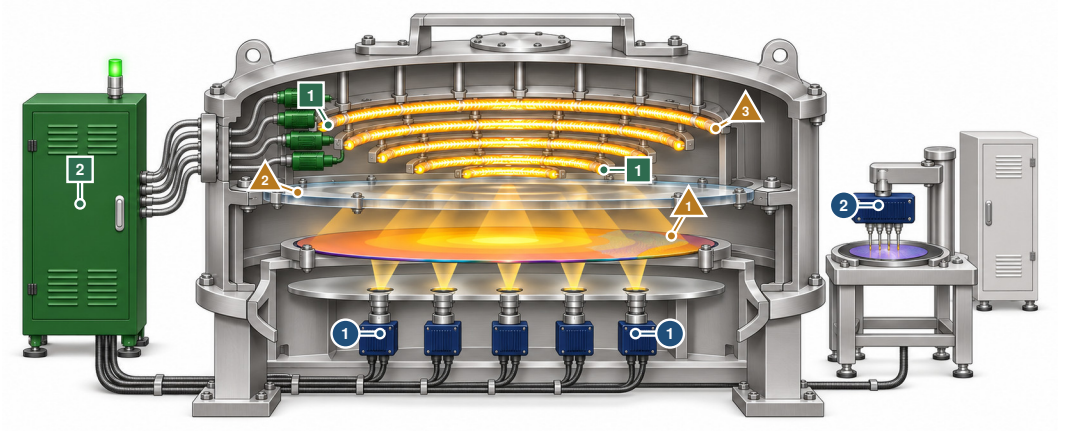
A process engineer asks why the fast loop moves cathode-bias current when the extraction voltage acts on the gap directly and the filament is the heater. Answer from the chapter's numbers for the nominal recipe.

8.12 Review the agent · 5 min

An AI assistant reviewing scenario 6 writes: "The 14.75 % overshoot is integrator windup. Raise the Back-calculation gain K_{aw} from 5 to 20/s so the integrator unwinds four times faster, and the overshoot will fall below 2 %, into the green band of the Overshoot tile." Find the flaw, and say how you would check the claim in the app before anyone changes a setting on a tool.

CHAPTER 9

Rapid Thermal Processing Simulator for Ramp Feedforward and Pyrometer Emissivity Error



The controller moves 1 the power of each of the five rings of lamps
2 the soak time, set from the power cabinet and corrected from wafer to wafer

The controller reads 1 five thermometers under the wafer that judge its heat by its glow
2 a probe that measures the finished wafer's sheet resistance

What disturbs it 1 a film on the wafer that makes it glow brighter than the thermometers assume
2 a quartz window that warms across a lot when its cooling water is turned down
3 lamps that age or fail, so a ring delivers less than it is told

Notice that the thermometers under the wafer read only its glow: a film like the patch at the right makes the wafer look hotter than it is, and only the probe station measures what the anneal did. Illustration, not to scale.

IN PLAIN TERMS Lamps That Bake a Wafer in Seconds

The main idea. A chip factory heats each thin silicon disk with rings of bright lamps, from warm to glowing hot and back in a few seconds. The heat wakes up the atoms planted in the silicon, and controllers steer the lamps from light sensors under the disk.

The challenge. The sensors judge heat by how brightly the disk glows. A thin coating can make a disk glow brighter without being hotter, so every reading looks perfect while the disk stays too cool. The lamp rings also light each other's parts of the disk.

Why it matters. Only a test of the finished disk shows whether the heat did its job.

THREE WAYS TO READ THIS CHAPTER

Explore	12 min	The first page, the history boxes of Section 9.2, the fact boxes of Sections 9.4 and 9.6, the try-it boxes of Section 9.6, and the key points
Study	55 min	The whole chapter with its pause points and the twelve exercises
Practice	20 min	The first page, Section 9.6, Section 9.7, Section 9.8 with its open challenge

The control question How can five lamp zones follow a spike anneal at 250 °C/s when the wafer's gain moves with the cube of its temperature, the zones light each other's pyrometers and the pyrometers may be converting radiance with the wrong emissivity? And what does a loop on its setpoint prove about the wafer it anneals?

Learning objectives

- Derive the radiative conductance and time constant of a wafer and explain why loops that ask for a heating rate need no gain schedule.
- Compute the power a ramp needs and the lag of a feedback loop without feedforward, and compare both with the inverse-model feedforward.
- Predict the temperature a loop holds when the emissivity is wrong, and name the measurements that reveal and correct the error.
- Relate a kelvin of peak temperature to the thermal budget, the sheet resistance and the soak correction between wafers.
- Compare a five-zone MPC with decoupled zone loops and a slip-aware ramp limit on a ramp too fast for the edge lamps.

Prerequisites Sections 1.6.2, 1.4.5, 1.4.2 and 1.4.3 of Chapter 1, and Section 1.6.3 for the MPC; Chapter 5 for multizone thermal control.

The app https://apmonitor.com/apps/rtp-wafer-anneal/

See also Volume 3, *Feedforward, Cascade, and Ratio Control*: Volume 3, Chapter 6 and Volume 3, Chapter 12 (a model feedforward that leaves feedback a trim). Volume 5, *Delays, Slow Measurements, and State Estimation*: Volume 5, Chapter 11 (a product temperature no instrument reaches). Volume 6, *Multivariable and Plantwide Control*: Volume 6, Chapter 2 (interacting loops and a static feedforward) and Volume 6, Chapter 3 (a sheet property measured after the step).

9.1 Introduction

An implanted wafer does not yet hold working transistors. The dopant atoms that will form each source and drain sit in silicon that the implant has damaged, and they become electrically active only when the wafer is heated, while the same heat lets them diffuse and deepens the junction. Rapid thermal processing (RTP), which heats one wafer at a time with lamps, offered an alternative to the furnace tube for this step [1, 2], and the anneal was later cut down to a spike, a ramp up and straight down, whose ramp-up rate is a process variable of its own [3]. The implant itself is the subject of Chapter 8. The chamber of this chapter takes a 300 mm wafer 775 µm thick from 600 to 1050 °C at 250 °C/s, holds it for 1.0 s and lets it cool. Near the top of that ramp the heat capacity alone asks for 31.5 kW, and the five lamp zones draw 106.8 kW of electricity to put 47.5 kW into the wafer.

Two properties of this plant are simple to state and hard to feel. The first is that its loss is radiative. The slope of the loss, the conductance that sets how quickly the wafer answers its lamps, rises from 8.2 W/K at 400 °C to 53.1 W/K at 1050 °C while the heat capacity rises only from 111.9 to 126.0 J/K, and the open-loop time constant drops from 13.7 to 2.4 s as a result. The second is that the temperature a loop regulates is a pyrometer's reading, computed from radiance with an emissivity that somebody supplied. The emissivity of silicon changes with temperature [4], and film stacks and patterns change it again [5]. In this model a nitride cap that the pyrometers take for bare silicon leaves them peaking at 1048.4 °C while the true center peaks at 1017.0 °C and the sheet resistance comes out 26.3 % high. The belief the simulator is built to dislodge is that a loop holding its setpoint proves the product was made. A second belief falls with it, that a gain moving with the cube of temperature must be scheduled, because loops that ask for a heating rate overshoot a 10 K step by 3.5 K at 600 °C and by 3.3 K at 1050 °C with no schedule at all.

None of this is peculiar to one chamber. RTP temperature control is a distributed thermal problem with a stringent uniformity, fast ramps and a thermal budget that shrinks as devices shrink, and the physics-based design route worked out for it applies equally to slower hot-wall batch furnaces [6].

Model-based control has been pursued since low-order models of the chamber's dynamics were identified [7], because it offers the greatest potential for tight control of the wafer temperature [8, 9]. Microelectronics manufacturing adds a product that is measured only after the step, so a slower correction between runs sits above the fast loops [10], and a run-by-run controller combining statistical process control with feedback was demonstrated on a silicon epitaxy reactor [11]. The chamber brings together five pieces that recur across this class of thermal processes, a gain that moves with temperature, a trajectory faster than feedback, heaters that light each other's targets, a non-contact sensor whose error depends on the material, and a quality read afterwards. A transfer function identified at one temperature carries none of them.

However, the published work on RTP control assumes a chamber in hand, and the standard course treatment linearizes a nonlinear plant about one operating point [12], so a student meets neither a loop that holds its reading while the wafer misses its budget nor the radiation that makes the ramp hard. This chapter works with a spike anneal chamber that runs in a browser, in which the true annulus temperatures are logged apart from the five pyrometers the controller reads, the product is an Arrhenius budget integrated along each wafer's own trajectory, and the lamps, the edge ring and the window each carry their own dynamics. The radiative model, the inverse-model feedforward, the decoupler, the correction between wafers and the model predictive controller all follow published practice. The model has 29 states with closed energy ledgers, a pyrometer chain that stray light and a wrong emissivity can fool, a supervisor that refuses a lot it cannot make and a five-zone model predictive controller measured against the loops it replaces, all written out in enough detail to rebuild the application. Nine cases and a graded lot pull apart a time constant and a gain, a ramp and its error, one zone and its neighbors, and a reading and the wafer.

Section 9.2 traces lamp heating, pyrometry and model-based control of the chamber. Section 9.3 sets out the chamber as modeled, from the annuli of the wafer to the thermal budget and the slip limit. Section 9.4 derives what radiation does to the loop: a time constant that falls with temperature, the power a ramp needs, zones that light each other and an error the loop cannot see. Section 9.5 describes the zone loops, the inverse-model feedforward, the supervisor, the emissivity modes, the soak correction and the model predictive controller. Section 9.6 reports what each case measures, the graded lot and the comparison of the predictive controller with the zone loops. Section 9.7 describes the real thermometers, lamps and controllers behind the loops, and Sections 9.8 and 9.9 give the limitations and the conclusions.

9.2 Lamps, Pyrometers and the Spike Anneal

Lamp heating entered the fab as an alternative to the furnace [1, 2], as the box below tells.

> ⧖ HISTORY **From the Furnace Tube to the Lamp (1988 to 1990)**
>
> Wafers were once heated in batches, in furnace tubes. In 1988 a review described rapid isothermal processing, heating one wafer with incoherent light, as a superior alternative to furnace processing. Its attractions were low cost, a minimum overall thermal budget, low power consumption and high throughput, and the review named more accurate temperature measurement as a development still needed [1]. Two years later a review written with its emphasis on temperature control summarized the options of 15 equipment makers, with the techniques of the day for measuring and controlling the wafer temperature and new optical, fiber-optic and photoacoustic ones [2]. In 2006 a dissertation still called the difficulty of measuring the wafer's temperature the biggest obstacle to the technology, which had kept its advantage over batch processing [13].

The rim was an early difficulty. A wafer lit uniformly loses more heat at its edge than in its middle, and a calculation of the temperature and stress fields in 1984 showed that the thermoelastic stress from that edge drop is much higher under uniform irradiation than under a strip heater. The same study resolved the stress onto the slip planes of the crystal and found the onset of damage on 2-inch wafers held for 15 s between 1030 and 1090 °C [14]. Heat transfer models with radiation and

convection were soon compared against the yield stress of silicon at the temperature and strain rate of the process, together with methods for improving the uniformity [15]. A lamp head divided into radial zones, with an outer zone that lights both the wafer's rim and the ring that supports it, is the arrangement the simulator carries.

Measurement was the second difficulty, because a pyrometer reads radiance and not temperature. The total hemispherical emissivity of lightly doped silicon 390 µm thick was measured to rise from 0.12 at 280 °C to a limiting value of 0.7 at 650 °C [4], and electromagnetic multilayer calculations coupled to a reactor model showed that film patterns change a wafer's radiative properties enough to matter for the temperature uniformity of shallow-junction anneals [5]. An emissivity assumed for one film stack can therefore be wrong for the next, and the error enters the loop through its only measurement.

Control followed the measurement into models [6–9, 16], as the box at the end of this section tells. Between runs, the run-by-run controller joined statistical process control to feedback and recovered an epitaxy process within 3 runs of a disturbance [11], and the practice of automatic control across microelectronics manufacturing was surveyed at length [10]. Spike annealing for ultra-shallow junctions was studied with the ramp-up rate as a variable of its own [3], and it is the recipe the simulator runs.

> ⧖ HISTORY **From Lamp Recipes to Model-Based Control (1993 to 2005)**
>
> In 1993 a dynamic model of a circularly symmetric chamber with several lamp zones, carrying radiation, conduction, convection and the heating of the chamber itself, was paired with a multivariable nonlinear model predictive controller. The study asked how the choice of the objective function changed the control of such a distributed system [16]. In 1997 an experimental chamber ran on a state-space model identified from data, a temperature trajectory shaped by convex optimization so that high-gain feedback could respect the lamps' limits, and linear quadratic Gaussian feedback [8]. By 2003 physics-based models, reduced to low order, ran as real-time embedded controllers on commercial chambers [6]. Yet in 2005 many tools still held one pyrometer at the wafer's center with one loop, the zone powers fixed as shares stored in the recipe [17].

9.3 The Chamber as Modeled

The simulated tool anneals a lot of 25 wafers, one every 60 s. Each slot loads a wafer at 25 °C with the lamps idle, preheats it open loop until all five pyrometers read, stabilizes it at 600 °C, ramps it at 250 °C/s to 1050 °C, soaks it for 1.0 s, lets it cool to 500 °C and unloads it. Figure 9.1 draws a half section of the chamber with the symbols of the equations below, and Table 9.1 classifies each variable as a controlled, manipulated, disturbance or state variable.

The 29 states are the eight annulus temperatures, the edge ring, the five delivered lamp powers, the window, eight thermal budgets and six energy ledgers of the wafer's flows and the lamps' electricity. Temperatures are in degrees Celsius on the panels and in kelvin inside every radiative and Arrhenius term.

9.3.1 The Wafer

The wafer is eight annuli of equal width $\Delta r = 18.75\,\text{mm}$ out to $R_w = 150\,\text{mm}$, each with one temperature. Annulus j balances the lamp power it absorbs against radiation and convection from both faces and conduction to its neighbors, as shown in Equation (9.1).

$$\begin{aligned} m_j c_p(T_j)\frac{\mathrm{d}T_j}{\mathrm{d}t} = &\sum_{k=1}^{5} G_{jk}P_k - \varepsilon(T_j)\,\sigma\,2A_j\,(T_j^4 - T_{ch}^4) \\ &- h\,2A_j\,(T_j - T_{gas}) + K_{j-1,j}(T_{j-1} - T_j) \\ &- K_{j,j+1}(T_j - T_{j+1}) + \delta_{j7}\,q_{rim} \end{aligned} \tag{9.1}$$

where $A_j = \pi(r_{j+1}^2 - r_j^2)$ with $r_j = j\Delta r$, $m_j = \rho\, t\, A_j$ with $\rho = 2329\,\text{kg/m}^3$ and $t = 775\,\mu\text{m}$, σ is the Stefan–Boltzmann constant, $T_{gas} = 25\,^\circ\text{C}$, $h = h_0 f_g^{0.6}$ with $h_0 = 8\,\text{W/(m}^2\,\text{K)}$ and f_g the gas flow as a

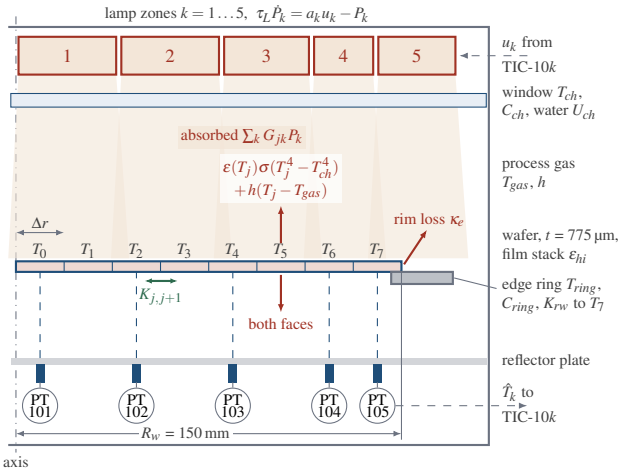


Figure 9.1 Half section of the lamp-heated chamber, drawn to scale in radius. Five lamp zones deliver P_k through a quartz window at T_{ch}, the eight annuli at T_0 to T_7 absorb $\sum_k G_{jk} P_k$, radiate and convect from both faces and conduct through $K_{j,j+1}$, and the rim loses heat through κ_e and exchanges it with the SiC edge ring at T_{ring}. Pyrometers PT-101 to PT-105 look up through the reflector plate at radii r_k and report $\hat{T}_k$ to the zone loops, which return the lamp commands u_k.

Table 9.1 Variables of the Chamber by Role, With Nominal Values of the Reference Configuration

Variable	Role	Nominal	Range
$\hat{T}_k, k = 1 \ldots 5$	CV, PT-101 to 105	recipe SP	350–1250 °C
T_{pre}, T_{peak}	recipe setpoints	600, 1050 °C	450–800, 950–1150
r, t_{soak}	recipe rate, soak	250 °C/s, 1.0 s	50–400, 0–4
R_s	quality, after the wafer	400 Ω/sq	±2 % band
ΔT	constraint, true	$< \Delta T_{slip}$	—
u_k	MVs, lamp commands	feedforward	0–100 % of rating
δ	MV, soak correction	0 s	−0.9 to 2.0 s
ε_{hi}	DV, film stack	0.70	0.63, 0.70, 0.77
a_3, a_5	DVs, lamp output	100 %	30–100
f_w, f_g	DVs, water, gas flow	100 %	10–150, 20–200
κ_e	DV, rim loss factor	1	0–3
σ_p	DV, pyrometer noise	0.3 K	0–3
T_j, $j = 0 \ldots 7$	states, annuli	25 °C at load	—
T_{ring}, T_{ch}	states, ring, window	281.5, 30 °C at reset	—
P_k	states, delivered power	1 % idle	—
θ_j	states, budgets	0 at load	—
$E_{abs}, \ldots$	states, six ledgers	0 at load	—

fraction of design, and the conductance between annuli is $K_{j,j+1} = k(\bar{T})\, 2\pi r_{j+1} t / \Delta r$ at the mean of the two temperatures. The terms at the rim, q_{rim}, follow in Equation (9.3). Equation (9.2) supplies the properties of silicon.

$$\begin{aligned} c_p(T) &= \left(a_0 + a_1\tau + a_2\tau^2 + a_3\tau^3 + a_4\tau^{-2}\right)/M_{Si} \\ k(T) &= \left(-4.55\times10^{-3} + 3.74\times10^{-5}\,T\right)^{-1} \\ \varepsilon(T) &= \varepsilon_{hi} - \frac{\varepsilon_{hi} - 0.30}{1 + \exp[(T - 430\,°\mathrm{C})/55\,\mathrm{K}]} \end{aligned} \tag{9.2}$$

The heat capacity is the NIST-JANAF Shomate fit for crystalline silicon in $\tau = T/1000\,\mathrm{K}$ with $a_0 = 22.81719$, $a_1 = 3.899510$, $a_2 = -0.082885$, $a_3 = 0.042111$ and $a_4 = -0.354063$ in J/(mol K) and $M_{Si} = 0.028\,085\,5\,\mathrm{kg/mol}$, which gives the whole wafer 90.8 J/K at 25 °C, 116.9 J/K at 600 °C

and 126.0 J/K at 1050 °C. The conductivity follows the phonon conductivity of silicon, 150 W/(m K) at 300 K and 24.8 W/(m K) at 1200 K, with T in kelvin. The total emissivity rises from 0.30 when cold, because lightly doped silicon is partly transparent in the infrared, to the film stack's high-temperature value ε_{hi} above about 600 °C, a shape taken from the measured rise of silicon's emissivity [4]. The three stacks are teaching values, 0.70 for bare silicon, 0.63 under a thermal oxide and 0.77 under a nitride cap, and each wafer draws its own value with a scatter of 0.1 %.

9.3.2 The Rim and the Edge Ring

The outer annulus carries two more terms. It loses heat through its edge face and through its weaker view of the reflector, lumped as an equivalent band $w_E = 3\,\text{mm}$ wide, and it exchanges radiation with a silicon carbide ring that supports it, as shown in Equation (9.3).

$$\begin{aligned} q_{rim} &= K_{rw}\,\sigma\,(T_{ring}^4 - T_7^4) \\ &\quad - \kappa_e\,\varepsilon(T_7)\,\sigma A_E\,(T_7^4 - T_{ch}^4) \\ C_{ring}\frac{\mathrm{d}T_{ring}}{\mathrm{d}t} &= \sum_k G_{rk}P_k - \varepsilon_r\sigma A_r(T_{ring}^4 - T_{ch}^4) \\ &\quad - K_{rw}\,\sigma\,(T_{ring}^4 - T_7^4) \end{aligned} \tag{9.3}$$

where $A_E = 2\pi R_w w_E = 2.83\times10^{-3}\,\text{m}^2$, κ_e is the rim loss factor, $K_{rw} = 2.8\times10^{-3}\,\text{m}^2$ is the exchange area between the ring and the rim, and the ring runs from 146 to 170 mm at a thickness of 0.7 mm with a density of 3210 kg/m^3 and a heat capacity of 1100 J/(kg K), so $C_{ring} = 58.9\,\text{J/K}$. It radiates from $A_r = 0.0439\,\text{m}^2$ with $\varepsilon_r = 0.85$ and absorbs 85 % of the lamp light that lands on it.

9.3.3 Lamps and the Gain Matrix

Each zone follows its command through a first-order filament lag, and each lights a blurred annulus of the wafer plane, as shown in Equation (9.4).

$$\begin{aligned} \tau_L\frac{\mathrm{d}P_k}{\mathrm{d}t} &= a_k u_k - P_k \\ G_{jk} &= \alpha_w\,\eta\,\frac{\int_{r_j}^{r_{j+1}}\phi_k(r)\,2\pi r\,dr}{\int_0^{R_p}\phi_k(r)\,2\pi r\,dr} \\ \phi_k(r) &= \tfrac{1}{2}\big[F_k(r) + F_k(-r)\big] \\ F_k(x) &= \operatorname{erf}\big[(x - b_{k-1})/s\big] - \operatorname{erf}\big[(x - b_k)/s\big] \end{aligned} \tag{9.4}$$

where the command u_k is in watts, a_k is the zone's output as a fraction of new, $\tau_L = 0.20\,\text{s}$, the footprints end at b_k = 40, 80, 115, 140 and 172 mm with $b_0 = 0$, the blur is $s = 19\,\text{mm}$, the mirrored term $F_k(-r)$ makes zone 1 a disc, $\eta = 0.85$ of a zone's power reaches the plane out to $R_p = 185\,\text{mm}$, and the wafer absorbs $\alpha_w = 0.70$ of what lands on it. The ring's row G_{rk} is the same integral from R_w to 170 mm with absorptivity 0.85. The ratings are 9, 22, 32, 30 and 52 kW, 145 kW in all, and an idle filament holds 1 % of each rating.

9.3.4 The Window and the Pyrometers

The quartz window is one lumped temperature. It absorbs a share of the lamp power and of the radiation leaving the wafer and the ring, and water cools it, which Equation (9.5) expresses.

$$\begin{aligned} C_{ch}\frac{\mathrm{d}T_{ch}}{\mathrm{d}t} &= \beta_L\sum_k P_k + \beta_W(q_{rad} + q_{edge} + q_{ring}) \\ &\quad - U_{ch}f_w(T_{ch} - T_{cool}) \end{aligned} \tag{9.5}$$

where $C_{ch} = 6\,\text{kJ/K}$, $\beta_L = 0.06$, $\beta_W = 0.10$, q_{rad}, q_{edge} and q_{ring} are the wafer's face radiation, its rim loss and the ring's radiation, $U_{ch} = 12\,\text{W/K}$, f_w is the water flow as a fraction of design with a

floor of 0.05, and $T_{cool} = 25\,^{\circ}\mathrm{C}$. Five pyrometers look at the back of the wafer at r_k = 9.4, 46.9, 84.4, 121.9 and 140.6 mm, the centers of annuli 0, 2, 4, 6 and 7, so annuli 1, 3 and 5 have no pyrometer of their own. Equation (9.6) gives what each one measures and how it converts the radiance.

$$\begin{aligned} \frac{R_k}{\sigma} &= \varepsilon(T)\,T^4 + f_s\,(T_{ch}^4 - T_{cal}^4), \qquad T_{bb,k} = (R_k/\sigma)^{1/4} \\ \hat{T}_k &= \left(T_{bb,k}^4/\varepsilon_{set}\right)^{1/4} + \sigma_p\, n_k \approx T\,(\varepsilon/\varepsilon_{set})^{1/4} \end{aligned} \tag{9.6}$$

where T is the true temperature at r_k, $f_s = 1$ scales an illustrative stray radiance from the window beyond its value at the 30 °C calibration, so the approximation holds while the window is at that temperature, $T_{bb,k}$ passes a first-order lag of 5 ms, n_k is unit colored noise with a 4 ms correlation time, and ε_{set} comes from the emissivity mode of Section 9.5. With the film-stack table the conversion solves $\varepsilon(\hat{T})\hat{T}^4 = T_{bb,k}^4$ with the table's own temperature dependence. A reading outside 350 to 1250 °C is flagged out of range and clamped.

9.3.5 Thermal Budget, Sheet Resistance and Slip

Dopant activation is represented by a thermal budget per annulus, counted in equivalent seconds at $T_n = 1050\,^{\circ}\mathrm{C}$, and a sheet resistance that falls as a power of it, as shown in Equation (9.7).

$$\frac{\mathrm{d}\theta_j}{\mathrm{d}t} = \exp\Big[-\frac{E_a}{k_B}\Big(\frac{1}{T_j} - \frac{1}{T_n}\Big)\Big], \qquad R_{s,j} = R_{s,T}\Big(\frac{\theta_T}{\theta_j}\Big)^m \tag{9.7}$$

where $E_a = 3.5\,\mathrm{eV}$, k_B is the Boltzmann constant, $R_{s,T} = 400\,\Omega/\mathrm{sq}$, $m = 0.30$, and $\theta_T = 1.3365\,\mathrm{s}$ is the budget of the nominal spike followed exactly by a uniform wafer, so that the nominal recipe makes the target by construction. The wafer's sheet resistance is the area mean of $R_{s,j}$, its range is taken at the eight annulus centers, and the four-point probe reports the mean with a repeatability of 0.3 %. A wafer is in specification when its true mean lies within 2 % of the target and it has not slipped. Slip is judged on the true temperatures at the eight annulus centers, as shown in Equation (9.8).

$$\begin{aligned} \Delta T &= \max_i T_i - \min_i T_i \;>\; \Delta T_{slip}(T_{max}) \\ \Delta T_{slip}(T) &= 12\,\mathrm{K}\cdot\exp\left[(1000\,^{\circ}\mathrm{C} - T)/110\,\mathrm{K}\right] \end{aligned} \tag{9.8}$$

held for 20 ms, where T_{max} is the hottest of the eight. The limit is 29.8 K at 900 °C, 12 K at 1000 °C and 7.6 K at 1050 °C. The budget, the power law and the slip limit are illustrative teaching values that rank thermal histories. They are not models of activation, diffusion or dislocation, and real slip limits come from comparing the thermal stress resolved on the slip planes with the yield stress of silicon at the temperature and strain rate of the process [14, 15].

9.3.6 Constants, Initial State and Numerics

Table 9.2 collects the constants. A visible constant can be moved with its slider in the Show Model panel (Section 1.6.4), a hidden one is not offered there, and a fixed one is written into the source. A lot opens with every annulus at 25 °C, the window at 30 °C, the lamps idle and the ring at its idle equilibrium of 281.5 °C, and each new wafer arrives at 25 °C with its budgets at zero. The held wafer of case 1 opens at the steady state of its lamp stations, solved by Newton's method on the nine wafer and ring balances with the window iterated to its own balance.

All 29 states are integrated together with a fourth-order Runge-Kutta scheme on a fixed 1 ms step, the controller runs on a 10 ms scan, and 20 rows a second reach the historian. The fastest state is the filament, with $\Delta t/\tau_L = 0.005$. Because the ledgers are integrated with the states, the wafer's stored enthalpy and its integrated flows agree over a whole wafer to 2.7×10^{-13} of the energy crossing its faces, and over that wafer it absorbs 213.4 kJ, radiates 145.3 kJ, convects 29.6 kJ and loses 2.9 kJ at the rim. The application's own checks set the limits for refinement. Sixteen annuli instead of eight move the first wafer of case 3, with the decoupler and the feedforward, by 0.16 % of sheet

Table 9.2 Constants of the Model and How the Interface Exposes Each One

Symbol	Quantity	Value	Access
R_w, t	wafer radius, thickness	150 mm, 775 µm	fixed
Δr, ρ	annulus width, density	18.75 mm, 2329 kg/m^3	fixed
ε_{hi}	bare, oxide, nitride	0.70, 0.63, 0.77	fixed
—	emissivity scatter, 1σ	0.1 %	visible
h_0	convection at design flow	8 W/(m^2 K)	visible
w_E, T_{gas}	rim band, gas	3 mm, 25 °C	fixed
τ_L	filament lag	0.20 s	visible
$P_{max,k}$	zone ratings	9, 22, 32, 30, 52 kW	visible
—	idle filament power	1 %	hidden
s, η, α_w	blur, reach, absorption	19 mm, 0.85, 0.70	fixed
C_{ring}, ε_r	edge ring	58.9 J/K, 0.85	fixed
K_{rw}	ring to rim exchange	2.8×10^{-3} m^2	fixed
C_{ch}	window heat capacity	6 kJ/K	hidden
β_L, β_W	window absorption	0.06, 0.10	hidden
U_{ch}	window cooling	12 W/K	visible
f_s	stray radiance	1.0	visible
—	pyrometer lag, range top	5 ms, 1250 °C	hidden
—	pyrometer range bottom	350 °C	visible
E_a, m	budget, power law	3.5 eV, 0.30	visible
$R_{s,T}$	target and band	400 Ω/sq, ±2 %	visible
—	probe repeatability	0.3 %	hidden
—	slip limit at 1000 °C	12 K	visible
—	e-folding, dwell	110 K, 20 ms	hidden

resistance, 0.02 K of center peak and 0.19 K at the edge pyrometer's radius, and halving the step moves it by 4×10^{-13} of sheet resistance and 6×10^{-5} K of center peak, the finer sampling of the peak. The engine, its seed and its speeds work as in Section 1.6.1. Here the seed is 20260927, so a lot replays identically, and any speed between a quarter of real time and twenty times real time leaves the trajectory unchanged.

9.3.7 Assumptions

Eight annuli do not resolve the rim's own dip. The lateral diffusion length of silicon at 1050 °C, $\sqrt{kt/(8\varepsilon\sigma T^3)}$, is about 4.8 mm, a quarter of an annulus, which is why slip and uniformity are judged at the annulus centers out to 141 mm rather than at the rim itself.

Radiation is gray and band averaged. One total emissivity serves both faces, both faces see the window's temperature, and the reflector plate adds no cavity enhancement. The conversion $T(\varepsilon/\varepsilon_{set})^{1/4}$ follows from that assumption, and it overstates the error of a real narrow-band instrument, as Section 9.4 quantifies.

The wafer absorbs a fixed 70 % of the lamp light at every temperature and for every stack. A cold, lightly doped wafer is partly transparent beyond 1.1 µm and absorbs less, which here would only slow the open-loop preheat. The footprints, the gain matrix, the ratings and the filament lag are teaching values, and the delivered power is proportional to the command.

The ring and the window are single lumped temperatures, and the stray radiance of the window is an illustrative term. The chamber is axisymmetric and the wafer does not rotate, so azimuthal nonuniformity, wafer bow, the gas flow pattern and pattern-dependent absorption are all absent, although patterns are known to matter [5].

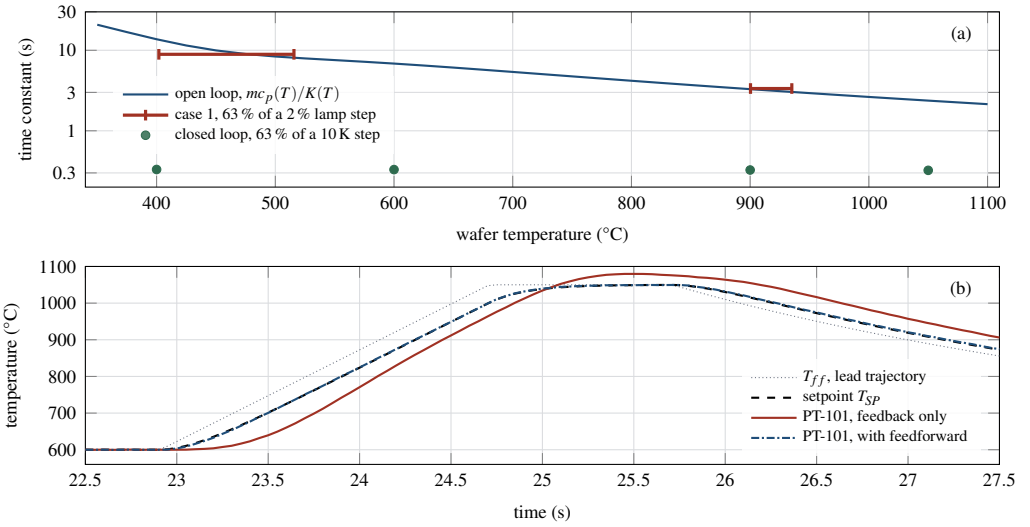


Figure 9.2 (a) Open-loop time constant of a uniform wafer from Equation (9.9) (solid line), with the 63 % times of the two lamp steps of case 1 drawn as bars across the temperatures each step spans and the 63 % times of a 10 K setpoint step under the zone loops as dots (pyrometer noise off). (b) First ramp of case 2 with one loop on the center pyrometer, as shipped (solid) and with the model feedforward (dash-dotted), against the setpoint (dashed) and the feedforward's lead trajectory (dotted).

9.4 What Radiation Does to the Loop

9.4.1 A Time Constant That Falls With Temperature

Linearizing Equation (9.1) for a uniform wafer yields the conductance $K(T)$, the slope of the loss with temperature, and with it the open-loop time constant of Equation (9.9).

$$\begin{aligned} K(T) &= \frac{\mathrm{d}}{\mathrm{d}T}\left[\varepsilon(T)\sigma 2A_w(T^4 - T_{ch}^4) + 2hA_w(T - T_{gas})\right] \\ \tau(T) &= m_w c_p(T)/K(T) \end{aligned} \tag{9.9}$$

where A_w and m_w are the area and mass of the whole wafer. With the chamber at 30 °C, K is 8.2, 17.1, 37.4 and 53.1 W/K at 400, 600, 900 and 1050 °C, and τ is 13.7, 6.8, 3.3 and 2.4 s, as shown in Figure 9.2. Above 800 °C the conductance is $4\varepsilon\sigma 2A_w T^3$ plus 1.13 W/K of convection. Below it the rising emissivity adds a slope of its own, so at 400 °C, where $\varepsilon = 0.447$, the fourth-power term alone gives 4.4 W/K and the emissivity's slope and the convection supply the rest. The heat capacity barely moves by comparison, from 111.9 to 126.0 J/K between 400 and 1050 °C, a factor of 1.13 against 6.5 for the conductance.

A plant gain that moves by 6.5 is the textbook argument for a gain schedule, the remedy Chapter 2 applies to a plant that moves across a whole run. The zone loops of Section 9.5 blunt it, because they ask for a heating rate and carry the heat capacity inside their own output, so their proportional action of 4 K/s per kelvin of error competes with a radiative rate $K/m_w c_p$ of only 0.15 to 0.42 s^{-1} between 600 and 1050 °C. Without pyrometer noise, the 63 % time of a 10 K setpoint step on the held wafer of case 1 in automatic is 0.33 s at every one of 400, 600, 900 and 1050 °C, as the dots in Figure 9.2(a) show.

🕮 **Pause.** Try Exercise 9.1 before reading on.

9.4.2 The Power a Ramp Needs

At 250 °C/s the heat capacity of 126.0 J/K asks for 31.5 kW near the peak, against the 17.15 kW that a uniform wafer at 1050 °C radiates into a 30 °C chamber and the 1.16 kW it convects. A feedback loop has to build that power from its own error, on a ramp that lasts 1.8 s. Less than half of the lamp power reaches the wafer. Zones 1 to 3 put 59.5 % of their power into the wafer, zone 4 54.4 % with 6.0 % into the ring, and zone 5 only 19.9 % into the wafer and 34.2 % into the ring. On the nominal ramp of the reference configuration the bank peaks at 106.8 kW electric for 47.5 kW absorbed, with the zones at 73 to 81 % of their ratings.

9.4.3 Zones That Light Each Other

Read at the five pyrometer spots, the gain matrix becomes a 5×5 matrix Φ of absorbed flux per watt, and it is far from diagonal. Section 1.4.5 treats the coupled zones of a heated plate in general, and Chapter 5 does so for a plate of a hundred zones. Per watt, the edge pyrometer receives 1.30 times more light from zone 4 than from its own zone 5, and the second pyrometer receives 1.19 times more from zone 1 than from zone 2. The ring, with a heat capacity of 58.9 J/K and nearly all of its light from zone 5, lags on a fast ramp, and the rim, radiating into it, lags with it. In case 3 with the decoupler and the feedforward, the ring is at 842 °C when the soak begins while the wafer averages 1001 °C, and still only at 914 °C when the soak ends.

9.4.4 An Error the Loop Cannot See

Equation (9.6) makes the pyrometer error multiplicative, as Equation (1.1) of Chapter 1 states in general. An emissivity 10 % higher than the pyrometer assumes makes it read $1.1^{1/4} = 1.024$ times the absolute temperature, so a loop holding the reading on 1050 °C holds a nitride-capped wafer at 1018.8 °C, 31.2 K lower. The same arithmetic runs the other way for an oxide stack at 0.63, whose wafer would be held near 1085.3 °C. A real pyrometer working in a narrow band at 0.95 µm follows Wien's law, $\Delta T/T = (\lambda T/c_2)\,\Delta\varepsilon/\varepsilon$ with $c_2 = 14\,388\,\mu\mathrm{m\,K}$, and makes an error of 0.87 %, or 11.6 K, about a third of the band-averaged model's, so the simulator overstates the effect on purpose. The window's own radiance adds a second, additive error once the chamber warms beyond its 30 °C calibration.

> ✦ DID YOU KNOW? **Two Degrees at a Thousand**
>
> In 2001 the roadmap of the chip industry asked the thermometers of rapid thermal tools for an accuracy better than 2 °C and a reproducibility better than 0.25 °C at 1000 °C [18]. In 2006 the best commercial systems near 1000 °C were estimated to be within 10 to 15 °C [13].

9.4.5 What One Kelvin Costs the Product

Near the peak the budget of Equation (9.7) changes by 2.3 % per kelvin and the sheet resistance by 0.70 % per kelvin, so the 2 % band on 400 Ω/sq is about 3 K of peak. Along the nominal setpoint, the 8 s at 600 °C contribute 1.1×10^{-6} equivalent seconds of the 1.3365 s budget, while 81.1 % of it accrues within 30 K of the peak, 61.1 % in the soak and 35.6 % in the cooling.

9.5 Implemented Controller

The controller reads only the five pyrometers and, on a calibration wafer, five thermocouples, besides the film stack that the lot's records declare. In each 1 ms step the order is the controller when its scan of $\Delta t_c = 10\,\mathrm{ms}$ is due, the integration, the slip test and wafer exchange, the outputs and the historian. The signal path of one zone is drawn in Figure 9.3.

9.5.1 Zone Loops That Ask for a Heating Rate

Each zone runs its own copy of the velocity-form PID of Section 1.6.2, TIC-101 to TIC-105, on its own pyrometer. Here it runs as a PI loop whose output is a heating rate v_k in kelvin per second, as

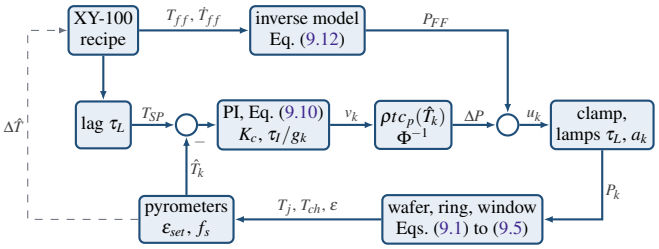


Figure 9.3 Signal path of one zone loop in the reference configuration. The supervisor's lead trajectory feeds the inverse model and, through the filament lag, the setpoint. The PI output is a heating rate that the areal heat capacity and the decoupler turn into lamp power, and the pyrometers' spread returns to the supervisor for the slip-aware ramp limit. Without the decoupler, Φ^{-1} becomes $1/\Phi_{kk}$.

shown in Equation (9.10).

$$\begin{aligned} e_n &= T_{SP,n} + \Delta T_k - \hat{T}_{k,n}, \qquad p_n = K_c e_n \\ S_n &= S_{n-1} + (p_n - p_{n-1}), \qquad w_n = S_n + I_{n-1} \\ v_n &= \mathrm{clamp}(w_n, v_{min}, v_{max}) \\ I_n &= I_{n-1} + \left[\frac{g_k K_c}{\tau_I} e_n + \frac{g_k (v_n - w_n)}{\tau_I} \right] \Delta t_c \end{aligned} \tag{9.10}$$

where ΔT_k is a zone offset from the recipe, zero by default, $K_c = 4\,\mathrm{K/s}$ per kelvin, $\tau_I = 1.5\,\mathrm{s}$, and $g_k = 1$ unless the schedule below is on. These are Equations (1.12) to (1.15) with p_n for P_k, w_n for v_k, v_n for u_k, no bias, no derivative and a tracking time equal to the integral time, so the sum S_n is never clamped and the back-calculation acts on the integral alone. There is no rate limit on the output, whose speed the filament lag already sets. A low reading raises its zone's heating, so each loop has action $a = +1$ and is reverse acting in the convention of Section 1.6.2. The app shows no action label.

The heating rate becomes lamp power through the controller's areal heat capacity and the gain matrix read at the spots, as shown in Equation (9.11).

$$\begin{aligned} \Delta f_k &= \rho\, t\, c_p(\hat{T}_k)\, v_k \\ \Delta P &= \Phi^{-1} \Delta f \quad \text{with the decoupler} \\ \Delta P_k &= \Delta f_k / \Phi_{kk} \quad \text{without it} \\ u_k &= \mathrm{clamp}\big(P_{FF,k} + \Delta P_k,\ P_{idle,k},\ P_{max,k}\big) \end{aligned} \tag{9.11}$$

where Δf_k is the absorbed flux the loop asks for at its spot and $P_{FF,k}$ is the feedforward below. With the decoupler all five zones together deliver exactly Δf at all five spots, the static decoupler of Section 1.4.5 written for absorbed flux. Without it, zone k alone answers loop k, and the loop's own limits v_{min} and v_{max} in Equation (9.10) are its lamp's idle floor and rating, less the feedforward, divided by $\rho t c_p(\hat{T}_k)/\Phi_{kk}$. With the decoupler the lamps are shared, the loops carry no limits of their own, and while any lamp is clamped each integral is driven back by $g_k(\bar{v}_k - v_k)\Delta t_c/\tau_I$ toward the output $\bar{v}_k$ that reproduces the lamps actually applied. The single-loop architecture of case 2 puts one loop on PT-101 and spreads its flux over the zones in the pattern that heats the whole wafer evenly. The idle floor gives way to zero in the cool phase. In MANUAL the operator sets the five commands, and every loop tracks by setting its integral so that its output reproduces the applied lamps, as it also does in the load, preheat and unload phases, so the return to automatic produces no bump, and a feedforward that appears or is switched is absorbed into the integrals in the same way.

The optional schedule multiplies each loop's integral gain by $g_k = [(\hat{T}_k + 273.15)/1323.15]^3$, the radiative conductance relative to its value at 1050 °C, so the integral time on the panel is the one at the peak.

9.5.2 The Inverse-Model Feedforward

The feedforward inverts a uniform-wafer version of Equation (9.1) along the recipe, as shown in Equation (9.12).

$$\begin{aligned} q_j &= m_j c_p(T_{ff})\,\dot{T}_{ff} + \varepsilon(T_{ff})\,\sigma\,2A_j\,(T_{ff}^4 - T_c^4) \\ &\quad + h_0\,2A_j\,(T_{ff} - T_{gas}) + \delta_{j7}\,\varepsilon\sigma A_E (T_{ff}^4 - T_c^4) \\ P_{FF} &= (G^\top W G)^{-1} G^\top W\, q, \qquad W = \mathrm{diag}(1/A_j) \end{aligned} \tag{9.12}$$

where the model assumes a cold chamber at $T_c = 30\,^\circ\mathrm{C}$ and the weighting makes the fit a least-squares match of absorbed flux over area. The filament lag is inverted by the reference itself. The supervisor keeps a lead temperature T_{ff} that moves at the recipe rate, computes the feedforward from it, and gives the loops the setpoint T_{SP} obtained by passing T_{ff} through the same lag, $\tau_L \dot{T}_{SP} = T_{ff} - T_{SP}$, which is what a wafer driven through real filaments by that feedforward can do. During cooling T_{ff} follows the model's free cooling of a uniform wafer with the lamps off. The model knows nothing of the edge ring, the warming window, the gas flow or a changed rim loss, and it uses the emissivity the pyrometers are set to. Its h_0, the lag τ_L of the reference and the exponent m of the soak correction below are the declared values of Table 9.2, so a learner who changes one of those constants changes a plant that the controllers do not know, and the mismatch reaches them only through the pyrometers and the probe.

9.5.3 Recipe Supervisor

XY-100 sequences each 60 s slot. It loads the wafer with the lamps idle for 6 s, preheats it open loop at 12 % of the bank in the uniform pattern until all five pyrometers read above 360 °C, and then hands over without a bump. The setpoint starts at the pyrometers' mean and continues at the heating rate the controller's model assigns to the preheat power, and each zone's setpoint starts from its own reading and joins the common one with a time constant of 0.8 s. The wafer is held at 600 °C for 8 s, ramped, soaked for 1.0 s from the moment T_{ff} reaches the peak, and cooled until T_{SP} reaches 500 °C. In case 3 with the decoupler and the feedforward, the phases change at 6.0, 10.3, 22.9, 24.7, 25.7 and 36.7 s. With the slip-aware limit on, the ramp rate is multiplied by a factor that follows, with a 20 ms lag, the target in Equation (9.13).

$$\rho_r = \mathrm{clamp}\Big(2\Big[1 - \frac{\max_k \hat{T}_k - \min_k \hat{T}_k}{\Delta T_{allow}(T_{peak})}\Big],\ 0.1,\ 1\Big) \tag{9.13}$$

where ΔT_{allow} is Equation (9.8) with the operator's 10 K at 1000 °C in place of the model's 12, evaluated at the peak, where the limit is tightest. The ramp therefore slows once the pyrometers' spread passes half of 6.3 K. A zone commanded to its rating for a total of 30 ms in the ramp or the soak marks the wafer lamp-limited, and, with the hold on, the supervisor stops the lot after that wafer rather than anneal wafers it cannot make.

9.5.4 Emissivity, Calibration and the Soak Correction

The emissivity the pyrometers use comes from one of three modes, a constant for every wafer, the film-stack table for the stack the lot's records say is loaded, or the last calibration wafer's estimate. A calibration wafer carries thermocouples at the five pyrometer radii, each lagging by 0.15 s with an offset of up to 0.5 K, and soaks for at least 3 s. Over the last 1.5 s of its soak each spot's estimate is the mean of $T_{bb,k}^4 / T_{tc,k}^4$, and the five are averaged. The estimate therefore absorbs the window's stray radiance as it was at calibration. The wafer-to-wafer correction adds δ to the soak after each wafer's probe reading, as shown in Equation (9.14).

$$\delta_n = \mathrm{clamp}\Big(\delta_{n-1} + \lambda\,\frac{\theta_T}{m}\ln\frac{\hat{R}_{s,n}}{R_{s,T}},\ -0.9\,\mathrm{s},\ 2.0\,\mathrm{s}\Big) \tag{9.14}$$

where $\lambda = 0.5$, because at the nominal spike one more second of soak changes $\ln R_s$ by $-m/\theta_T$. It is an integral controller between runs in the manner of run-by-run control [11], the incremental form of Equation (1.3) with the soak as the recipe and $\ln R_s$ as the result, which Chapter 6 applies to an etch trim. The soak rises when the sheet resistance is above target, so the correction is direct acting in the convention of Section 1.6.2.

9.5.5 Model Predictive Control of the Five Zones

The zone control may instead be a model predictive controller (MPC), the technology that carries most multivariable constraint control in the process industries [19], set up as Section 1.6.3 describes. It acts in AUTO through the stabilization, the ramp, the soak and the cool, and it replaces only the zone loops. The supervisor, its recipe, the lamp-limit flag and the hold are unchanged, the feedforward is not applied, and the five loops track the applied lamp commands so that they can take over without a bump. Its model is the zone loops' own, written at the five pyrometer spots with ten states, the spot temperatures $\tilde{T}_i$ and the delivered powers P_k, as shown in Equation (9.15).

$$\begin{aligned} c_a \frac{\mathrm{d}\tilde{T}_i}{\mathrm{d}t} &= \sum_{k=1}^{5} \Phi_{ik} P_k - L_i(T_i^\star) - \gamma_i(\tilde{T}_i - T_i^\star) + c_a d_i \\ \tau_L \frac{\mathrm{d}P_k}{\mathrm{d}t} &= u_k - P_k \\ L_i(T) &= (2+\chi_i)\,\varepsilon(T)\,\sigma\,(T^4 - T_c^4) + 2h_0(T - T_{gas}) \end{aligned} \tag{9.15}$$

where $T_i^\star(t)$ is the reference at spot i (the recipe setpoint plus the zone's offset), $c_a = \rho\, t\, c_p$ is the areal heat capacity at the mean of the five references at the moment of the decision, $\gamma_i = dL_i/dT$ at $T_i^\star$ at that moment, $\chi_i = 0$ except $\chi_5 = A_E/A_7 = 0.171$ for the rim's extra loss at the edge spot, $\varepsilon(T)$ is the emissivity the pyrometers are told, $T_c = 30\,^\circ\mathrm{C}$, h_0 and τ_L are the declared values, and d_i is an unmeasured disturbance in kelvin per second. The slope γ_i, the cubic gain of Section 9.4 per unit area, is 242 W/(m^2 K) at 600 °C and 752 at 1050 °C, and it and c_a are evaluated again at every decision, while the loss L_i itself enters along the reference over the horizon as a known time-varying term, so the plan knows that the top of the ramp costs more power than its start. The model has no radial conduction, no ring, no window and no aged lamps, and d_i absorbs them.

Every 50 ms, five scans, the controller discretizes Equation (9.15) for a zero-order hold on a 0.1 s grid and solves the quadratic program of Equation (9.16) over $N = 50$ steps, a 5 s horizon, with each command held over eight blocks, short near the present (1, 1, 2 and 2 steps) and long toward the end (4, 6, 10 and 24 steps), so the plan has 40 decision variables.

$$\begin{aligned} &\min \sum_{n=1}^{N} \Big[\sum_{i=1}^{5} e_{i,n}^2 + \sum_{i=1}^{4} (e_{i,n} - e_{i+1,n})^2 \Big] + w_u \sum_{\ell=1}^{8} \|\Delta u_\ell\|^2 \\ &\text{s.t. } \underline{u}_{k,\ell} \le u_{k,\ell} \le P_{max,k}, \qquad e_{i,n} = \tilde{T}_{i,n} - T_{i,n}^\star \end{aligned} \tag{9.16}$$

where both temperature weights are 1/K^2, the second sum is a weight on the difference between neighboring spots that makes the across-wafer spread count as much as tracking, $w_u = 0.04/\mathrm{kW}^2$ prices a 5 kW move per block like 1 K of error at one step, Δu_ℓ is the change of the five commands at the start of block ℓ (from the applied commands for the first block), and the floor $\underline{u}_{k,\ell}$ is the idle filament power for every block that starts before the preview reaches the cool phase and zero after it, as the loops clamp their own lamps. Every limit is hard and there is no soft constraint. The quadratic program is solved and applied in receding horizon as in Section 1.6.3 [20]. The preview is the supervisor's own reference over the horizon, replayed with the declared lag, the soak it will set, the ramp at the slip-aware factor of the moment and the free cooling after the soak. With the preview Off the reference is held at its present value, and the loops' feedforward receives the same recipe one lamp lag ahead.

Table 9.3 Controller Settings of the Reference Configuration, With the Ranges the Panel Allows

Setting	Value	Range
architecture	five zone loops	one loop on PT-101, MPC
MPC preview	on	off
K_c, τ_I at 1050 °C	4 K/s per K, 1.5 s	0–20, 0.1–20 s
feedforward, decoupler, schedule	on, on, off	on or off
T_{pre}, preheat power	600 °C, 12 %	450–800, 5–60
slip-aware limit, ΔT_{allow}	on, 10 K at 1000 °C	1–40 K
hold on a lamp limit	on	on or off
emissivity mode, constant	film-stack table, 0.70	constant, calibration
soak correction, λ	off, 0.5	0–1.5
scan, output, direction	10 ms, K/s, reverse	—

The disturbances come from a steady-state Kalman filter on each spot, run at every 10 ms scan. The lamp powers are propagated through the declared lag from the applied commands, and each pair $(\tilde{T}_i, d_i)$ is corrected by its pyrometer with the gain (0.0965, 0.317) of a Riccati solution for the spot model at 1050 °C, a pyrometer noise of 0.3 K and random walks of 0.02 K and 0.1 K/s per step on the temperature and the disturbance, so the estimation error decays with a time constant of 0.19 s. An input disturbance of this kind makes the controller offset-free at the five spots [21]. Acceptance of a plan requires an optimal status, a finite first move, plan, prediction and cost, and every planned command inside its block's limits. Otherwise the historian is marked, the five loops, which have been tracking the lamps, carry on from the applied commands, and the next accepted plan takes the lamps back over 0.3 s. With the noise off, the hand-over moves a lamp by at most 0.47 % of its rating, and a second rejection inside that return by at most 0.15 %.

Step tests check the model against the plant. On a held wafer, each zone stepped by 1 % of its rating for 20 s with the pyrometers held at 1050 °C gives a gain to its own spot within 6 % of the model's (−0.3 to +5.8 %) and a 63 % time of 2.60 to 2.69 s against 2.39 to 2.57 s. At 600 °C the model's gains are 15 to 26 % high and its 63 % times of 6.5 to 7.0 s are longer than the plant's 5.7 to 6.1 s, because conduction, which the model leaves out, spreads each zone's heat to the neighboring spots (zone 2 heats spot 1 by 20.0 K/kW against the model's 4.2). On the ramp of case 3 the model's 0.5 s prediction from the applied commands runs 2.9 K warm on average at the four inner spots (1.9 K with c_a taken along the path instead of at the decision), and in case 9 the aged edge lamps put the edge spot's prediction up to 21.6 K off until d_5 reaches about $-46\,\mathrm{K/s}$.

9.5.6 Three Models Kept Apart

Section 9.3 describes the plant that is integrated. What the controller carries inside is much smaller, namely the inverse model of Equation (9.12), the matrix Φ and its inverse, the areal heat capacity $\rho t c_p(\hat{T}_k)$, the free cooling of a uniform wafer and the soak sensitivity $-m/\theta_T$, and none of these contains the ring, the window or the true emissivity. The MPC's model of Equation (9.15) is assembled from the same pieces, and its estimator sees the plant only through the pyrometers and the applied commands. No identified control model exists in the application, and the open-loop steps of case 1 measure the plant the loops see, 63 % in 8.9 s near 400 °C and 3.4 s near 900 °C. Table 9.3 lists the shipped settings.

9.6 Measured Behavior of Each Case

Table 9.4 summarizes what each case changes and what it teaches. Every case loads the reference configuration of Table 9.3 and then applies only the changes in its row, and the outcomes come from the engine at its default seed with no interface attached, timed in simulated seconds. A variant is the case with one setting changed before the run starts, as each briefing asks after a Reset, which keeps the learner's settings (Section 1.6.4). A setting changed before Run and the same setting carried through a Reset give one run. Table 9.5 gives the first wafer of every single-wafer case and variant under the

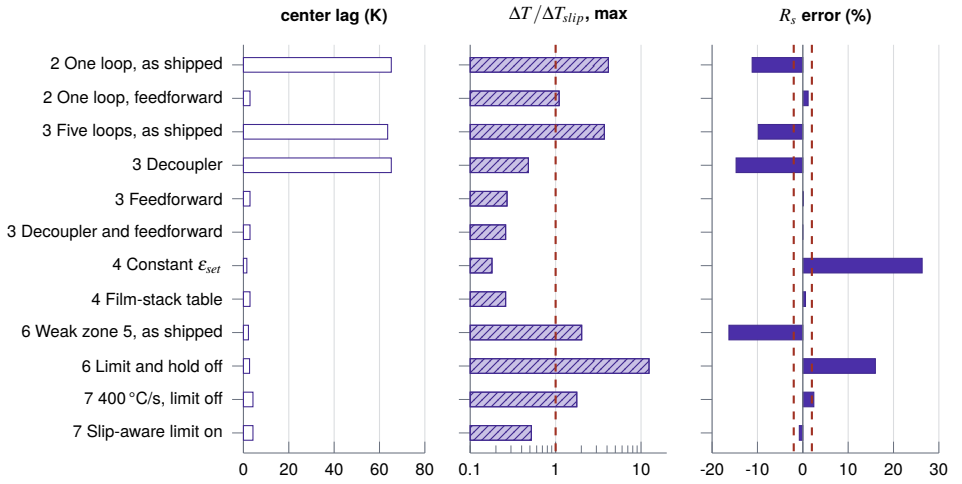


Figure 9.4 Three columns of Table 9.5 redrawn: the largest center lag on the ramp (open bars), the worst ratio of the true radial difference to the slip limit on a logarithmic axis (hatched, the slip limit of 1 dashed) and the sheet resistance error of the first wafer (solid, the 2 % band dashed). Notice that every row with the feedforward has a small lag and a sheet resistance inside the band, except case 4 with the constant emissivity.

Table 9.4 Scenario Sequence, From a Held Wafer to the Graded Lot

#	Scenario, changes from the reference	Teaching point
1	MANUAL, held wafer, +2 % at 400 and 900 °C	the cubic plant
2	one loop on PT-101, no feedforward[a]	a late ramp
3	five loops, no feedforward or decoupler[a]	loops that fight
4	nitride stack C, constant $\varepsilon_{set} = 0.70$	a trend that lies
5	stack C, calibration wafer first	an independent sensor
6	zone 5 at 35 % of its output	honestly infeasible
7	400 °C/s, zone 5 at 85 %[a]	rate against gradient
8	cooling water at 25 %, 25 wafers	a slow drift
9	case 7 under the MPC[a]	planning at the limits
10	constant ε, no feedforward or decoupler, stack C from wafer 10, zone 3 at 85 % from 16	the graded lot

[a] with the slip-aware ramp limit and the hold on a lamp limit off

zone loops, and Section 9.6.9 compares the MPC with them. Figure 9.4 draws three columns of that table as bars.

9.6.1 One Lamp Step at Two Temperatures

Case 1 holds one wafer in MANUAL with the stations at 1.7 to 1.8 %, 1.13 kW absorbed, where the center settles at 401.9 °C, the edge at 398.3 °C and the ring at 376.7 °C. At 5 s every station rises by 2 points, 1.28 kW more absorbed, and the center climbs 113.8 K by 45 s, 63 % of the way in 8.9 s. At 45 s the stations move to the feedforward's steady powers for 900 °C, 17.1 to 19.3 %, the wafer settles at 900.3 °C, and the same step at 75 s raises it only 34.9 K, 63 % of the way in 3.4 s. The conductance at 900 °C is 4.6 times that at 400 °C, so the same power buys 3.26 times less rise, 2.66 times faster, and the large step at 400 °C shortens its own time constant as it climbs, from 13.5 to 8.1 s between its two ends. In AUTO the loops hide most of this. On the same held wafer, a 10 K step of the hold setpoint overshoots on the center pyrometer by 3.5 K at 600 °C and 3.3 K at 1050 °C (3.1 and 2.9 K without pyrometer noise) and reaches 90 % in 0.45 s at both, and the schedule trims the overshoot at 600 °C to 2.2 K.

Table 9.5 First Wafer of Each Single-Wafer Case and Variant. Lags Are the Largest Setpoint Minus Pyrometer Reading on the Ramp, the Peak Is the True Center Peak, and ΔT Is the Largest True Radial Difference on the Ramp and Soak

#	Configuration	Ramp (s)	Center lag (K)	Edge lag (K)	Peak (°C)	ΔT (K)	$\Delta T/\Delta T_{slip}$ max	R_s error (%)	Range (%)
2	one loop, as shipped	1.80	65.2	77.6	1079.5	25.1	4.15	−11.17	17.80
2	one loop, feedforward	1.80	2.9	5.1	1049.1	8.4	1.10	+1.17	5.56
3	five loops, as shipped	1.80	63.6	41.4	1078.6	34.8	3.72	−9.81	7.87
3	decoupler	1.80	65.2	65.4	1079.5	2.9	0.48	−14.71	1.48
3	feedforward	1.80	2.9	2.6	1049.1	2.0	0.27	+0.06	1.41
3	decoupler and feedforward	1.80	2.9	2.9	1049.1	2.0	0.26	−0.05	1.40
4	constant ε_{set}, as shipped	1.80	1.5	1.5	1017.0	1.9	0.18	+26.31	1.32
4	film-stack table	1.80	2.9	3.0	1049.1	2.0	0.26	+0.59	1.37
6	weak zone 5, as shipped	9.61	2.2	12.4	1049.5	12.6	2.02	−16.29	1.57
6	limit and hold off	1.80	2.7	99.4	1049.3	104.9	12.41	+16.00	50.39
7	400 °C/s, limit off	1.13	4.2	15.6	1048.6	18.6	1.77	+2.43	1.48
7	slip-aware limit on	1.94	4.2	8.2	1048.6	6.3	0.52	−0.77	1.68

9.6.2 A Ramp Built From Its Own Error

Case 2 drives all five zones from one loop on PT-101 with no feedforward. On the ramp from 22.9 s the center reading falls up to 65.2 K behind its setpoint, because a proportional loop needs an error to ask for the ramp's power, and when the setpoint stops the integral that has finally caught up overshoots by 29.9 K, taking the true center to 1079.5 °C, as shown in Figure 9.2(b). The edge pyrometer lags by 77.6 K, the wafer slips at 24.7 s at 4.15 times the limit, and its sheet resistance is 355.3 Ω/sq, 11.2 % low, with a range of 17.8 % across the wafer. The feedforward supplies the ramp's power one lamp lag ahead, the center lag falls to 2.9 K with no overshoot and the sheet resistance is 404.7 Ω/sq. The edge still lags by 5.1 K and the wafer still slips, at 1.10 times the limit, because a single loop on the center cannot see the rim.

> **TRY IT IN THE APP** **2 · One Loop on the Center**
>
> Load the scenario and press Run. Watch the center pyrometer strip and the fast scope as the first ramp starts at about 23 s. The reading falls up to 65.2 K behind its setpoint and then overshoots by 29.9 K when the setpoint stops. At 60 s the Lot view gives the first wafer's sheet resistance, 355.3 Ω/sq, against a target of 400 Ω/sq.

9.6.3 Five Loops That Fight

Case 3 gives each zone its own loop without the feedforward or the decoupler. Each loop's correction also lands on its neighbors' spots, the middle loops get help while the center does not, and on the ramp the center lags 63.6 K while the edge lags only 41.4 K. The radial difference reaches 34.8 K and the wafer slips at 24.8 s, 3.72 times the limit. The decoupler spreads each correction through Φ^{-1} so that it moves only its own pyrometer. The wafer then lags by 65 K as one piece, with the radial difference at 2.1 K on the ramp and no slip, but it still overshoots by 29.9 K and comes out 14.7 % low. With the feedforward as well, the radial difference is 1.5 K on the ramp, the sheet resistance is 399.8 Ω/sq and its range 1.40 %. The feedforward alone does nearly as much, 1.6 K on the ramp and 0.06 % of error, because once the model supplies the power the loops have little left to fight over.

Pause. Try Exercise 9.3 before reading on.

9.6.4 A Perfect Trend on the Wrong Wafer

Case 4 loads nitride-capped wafers, emissivity 0.77, under pyrometers that assume 0.70, with everything else as the reference. The pyrometers read 2.4 % hot and the loops hold them on the recipe,

so the pyrometers peak at 1048.4 °C while the true center peaks at 1017.0 °C, 31.4 K lower. The first wafer's sheet resistance is 505.2 Ω/sq, 26.3 % high, and the second also 505.2 Ω/sq. Nothing on the loops' own trends shows it, as Figure 9.5 records during the first soak, and only the dashed true temperature, which no real tool has, and the probe after the wafer do. With the film-stack table the first wafer is 402.3 Ω/sq, 0.6 % high.

> **DID YOU KNOW? The Best Trend in the Table**
>
> From the simulator: in Table 9.5 the nitride-capped wafer read with the constant emissivity has the smallest lags of any row, 1.5 K at the center and at the edge, and the smallest radial difference, 1.9 K. Its sheet resistance is 26.31 % high, the largest error of any wafer in the table that did not slip.

Case 5 runs the same stack with the calibration mode, and the first wafer carries the thermocouples. Over the last 1.5 s of its 3 s soak the thermocouples read 31.8 K below the center pyrometer, which still assumes 0.70. The five estimates lie between 0.7705 and 0.7720 and average 0.7712, against the wafer's own 0.7706, and from the second wafer the pyrometers use them. Wafer 2 is 401.7 Ω/sq, 0.42 % high, and wafer 3 401.4 Ω/sq, 0.34 % high, where the wafers of case 4 were 26 % high. The estimate carries the window's stray light as it was at calibration, so it goes out of date as the chamber warms or the stack changes.

Pause. Try Exercise 9.2 before reading on.

9.6.5 An Edge Zone That Cannot Make the Recipe

Case 6 leaves zone 5 at 35 % of its output, 18.2 kW of its 52 kW rating, a third of which heats the ring. The zone sits at its rating for 1.36 s of the ramp and 0.48 s of the soak, so even the soak at 1050 °C needs more than it has. The slip-aware limit stretches the ramp from 1.80 to 9.61 s, the edge still falls 12.4 K behind, the wafer slips at 34.7 s at 2.02 times the limit, and the long slow ramp leaves its sheet resistance 16.3 % low. When it leaves at 60 s the supervisor holds the lot. With the limit and the hold off the edge pyrometer falls 99.4 K behind its setpoint at 250 °C/s, the radial difference reaches 104.9 K in the soak and the sheet resistance ranges over 50.4 % of its mean. Slower recipes do not rescue it. At 150, 100 and 50 °C/s with the limit off, zone 5 is at its rating through the whole soak and the wafer slips at 7.95, 4.51 and 1.95 times the limit. Some faults are not control problems, and this case is built to show one.

9.6.6 A Ramp Too Fast for the Rim

Case 7 ramps at 400 °C/s with zone 5 aged to 85 % and the slip-aware limit off. All five zones reach their ratings on the ramp, zone 5 for 0.55 s, the edge pyrometer lags by 15.6 K, the radial difference reaches 17.3 K on the ramp and 18.6 K in the soak, and the wafer slips at 24.2 s, 1.77 times the limit. With the limit on, the ramp slows while the spread is large and takes 1.94 s instead of 1.13, the edge lag stays at 8.2 K, the wafer reaches 0.52 of the limit and its sheet resistance is 396.9 Ω/sq, 0.8 % low, inside the band.

9.6.7 A Window That Warms Across the Lot

Case 8 runs 25 wafers with the cooling water at a quarter of its design flow. The window reaches 170.4 °C by the end of the lot, against 90.0 °C when the same lot runs at design flow with all 25 wafers in specification, and its radiance beyond the 30 °C calibration makes the pyrometers read high. The true center peak falls from 0.9 K below 1050 °C on the first wafer to 5.0 K below on the last, and the sheet resistance rises from −0.05 to +2.87 %. Wafer 20 is the first outside the band, wafer 22 falls back inside it at 1.94 %, and 20 of the 25 are in specification, against 18 to 20 at six other seeds of the random generator. With the wafer-to-wafer correction the soak grows to 1.10 s and all 25 are in specification, the worst 0.75 % off (Table 9.6). It repairs the product without knowing the cause, and the peak stays 5.0 K low.



Figure 9.5 The chamber in case 4, emissivity wrong by 10 %, at $t = 26.50$ s, every control as the case sets it, in the Chamber view with a 10.0 s trend window, cropped to the plant view and the trends. Wafer 1 of 25, nitride capped and read by pyrometers that assume bare silicon, is in the last tenth of its 1.00 s soak at a setpoint of 1049.4 °C. The cutaway (2) shows the five lamp zones, the window at 35 °C, the wafer, the edge ring at 897 °C and all five pyrometer beams, with the bank delivering 44.2 kW of which the wafer absorbs 19.0 kW. The disc (3) paints each annulus by its true difference from the wafer mean of 1016.2 °C, a spread of 1.8 K in which the annulus from 94 to 113 mm, which no pyrometer sees, runs 1.3 K cold. The badges (1) give the center pyrometer at 1048.2 °C, the edge 1.1 K below it, the peak so far at 1047.9 °C and no sheet resistance yet. The fast scope's (4) last 6 s run through the end of the stabilization, the ramp from 23.8 s and the soak, with both pyrometers on the setpoint throughout and the true center and edge 13.5 K below them at 600 °C and about 31 K below them in the soak. The trends (5) show zone 1 at 1048.2 °C with its true temperature of 1016.6 °C dashed and its lamps at 27.1 % beside a feedforward of 27.1 %, zone 5 at 1047.1 °C with 1016.1 °C true and its lamps at 34.6 % beside 30.7 %, and the true radial difference at 1.8 K, 1.6 K between the pyrometers, under a slip limit of 10.3 K.

9.6.8 The Graded Lot

The challenge runs 1500 s, with the nitride stack from wafer 10 and a zone 3 lamp string failing to 85 % at wafer 16. It ships with five independent loops, no feedforward, no decoupler and the constant emissivity, and it asks for all 25 wafers within 2 % of 400 Ω/sq, none slipped, and every true center peak within 3 K of 1050 °C, with the case's model parameters and disturbances unchanged. As shipped no wafer is in specification. Wafers 1 to 9 overshoot to true center peaks of 1059.1 to 1060.6 °C, 9.1 to 10.3 % low, and all slip. Wafers 10 to 15 peak at 1026.8 to 1028.3 °C while their pyrometers still reach about 1060 °C, 13.5 to 15.3 % high, and four of them slip. Wafers 16 to 25 peak at 1022.2 to 1023.0 °C, 5.0 to 8.1 % high, and all slip, 23 slips in all (21 to 23 at the six other seeds).

Table 9.6 Lots of 25 Wafers. The Peak Is the True Center Peak Less 1050 °C, and the Energy Is the Lamps' Electricity Over 1500 s

Configuration	In spec	Slips	R_s error (%)	Peak (K)	Energy (kWh)
case 8, cooling water at 25 %					
as shipped	20	0	−0.22 to +2.87	−4.95 to −0.60	3.29
soak correction	25	0	−0.65 to +0.75	−4.96 to −0.58	3.30
graded lot					
as shipped	0	23	−10.27 to +15.34	−27.8 to +10.6	4.06
shipped, table on	0	25	−14.66 to −7.70	+3.7 to +10.5	4.12
suggested[a]	25	0	−0.60 to +1.23	−2.32 to −0.55	3.42
with the schedule	25	0	−1.26 to +1.33	−2.29 to −0.63	3.45
no decoupler	25	0	−0.32 to +1.38	−2.35 to −0.56	3.41
no feedforward	0	0	−14.89 to −12.65	+24.8 to +29.8	3.50
constant ε	9	0	−0.14 to +27.26	−34.6 to −0.8	3.39
and soak correction	19	0	−0.58 to +26.67	−33.7 to −0.8	3.61

[a] film-stack table, feedforward and decoupler, and the rows below it each change one element of it

With the film-stack table, the feedforward and the decoupler, all 25 are in specification between −0.60 and +1.23 %, the worst wafer reaches 0.54 of the slip limit, the true center peak stays 2.32 to 0.55 K below 1050 °C as the window warms to 93.3 °C, and the lamps use 3.42 instead of 4.06 kW h. The worst sheet resistance leaves a margin of 0.77 points to the band and the worst peak a margin of 0.68 K. The same lot passes at the six other seeds, its worst wafer 1.13 to 1.44 % off.

The variants in Table 9.6 say which element carries the lot. Once the feedforward is on, the lot passes without the decoupler (worst 1.38 %, 0.55 of the slip limit) and with the schedule (1.33 %), and the differences among these three worst wafers lie within their spread over the six other seeds. Without the feedforward no wafer is in specification, every one overshooting by 25 to 30 K with the decoupler and all 25 slipping without it, and without the table the lot is lost from wafer 10. The soak correction does not replace the table. With the constant emissivity it lengthens the soak to 2.53 s and brings wafers 16 to 25 back inside the band, but wafers 10 to 15 are already lost and every nitride wafer still peaks about 33 K low.

9.6.9 The Predictive Controller Against the Zone Loops

Table 9.7 compares the first wafer under ten structures on two recipes, the nominal 250 °C/s ramp of case 3 and the 400 °C/s ramp of case 9 with zone 5 aged to 85 %, which is case 7 under the MPC. Every row shares the seed, the noise sequence, the pyrometers and the supervisor, whose recipe drives both the loops' feedforward and the MPC's preview. The MPC keeps the settings shipped in case 9 on both recipes. Row 4 retunes the loops once with the SIMC rule applied to the decoupled channel, an integrator behind the 0.2 s filament lag with $\tau_c = 0.2$ s, which gives $K_c = 2.5$ per second and $\tau_I = 1.6$ s, with the cubic schedule and the same feedforward. Rows 5 to 7 remove one mechanism from the MPC each, and rows 9 and 10 turn the slip-aware ramp limit on.

On case 9 the MPC holds the wafer where every loop structure without the slip-aware limit slips. Its largest radial difference on the ramp is 10.89 K, 0.79 of the limit, against 17.27 K and 1.77 times the limit for the decoupled loops with the feedforward, and with less than half of their lamp travel (685 against 1614 kW). A plan that cannot give the edge more power gives the inner zones less, as shown in Figure 9.6. The wafer pays for it in budget, because its sheet resistance is 3.34 % high, outside the band, against the loops' 2.43 % on a wafer that slipped. The retuned loops slip further (2.28 times the limit), so retuning the loops does not recover the edge.

Table 9.7 First Wafer Under Each Structure on Case 3's Recipe and on Case 9. ΔT Is the Largest True Radial Difference on the Ramp, the Overshoot Is the True Center Peak Less T_{peak}, Saturation Is the Time Any Zone Is Commanded at Its Rating in the Ramp and Soak, the Ratio Is the Worst $\Delta T/\Delta T_{slip}$, and the Travel Is the Summed Change of the Five Commands

		case 3 recipe, 250 °C/s					case 9, 400 °C/s, zone 5 at 85 %					
#	Structure	ΔT (K)	Overshoot (K)	R_s error (%)	Ratio	Travel (kW)	ΔT (K)	Overshoot (K)	Saturation (s)	R_s error (%)	Ratio	Travel (kW)
1	five loops, feedforward	1.63	−0.91	+0.06	0.27	1240	15.25	−1.40	0.49	+2.36	1.60	1248
2	and the decoupler	1.50	−0.91	−0.05	0.26	1633	17.27	−1.44	0.55	+2.43	1.77	1614
3	MPC	2.11	+0.82	+1.11	0.52	568	10.89	+0.42	0.75	+3.34	0.79	685
4	loops retuned once	1.46	−0.71	−0.07	0.25	1113	20.44	−0.92	0.43	+3.41	2.28	1131
5	MPC without lamp rows	2.07	−0.88	+0.41	0.29	537	11.69	−0.91	0.70	+2.51	0.84	607
6	MPC, preview off	2.11	−0.96	+0.89	0.45	544	6.76	−1.12	0.75	+3.35	0.76	602
7	MPC without spread weight	2.33	+0.85	+0.98	0.78	490	11.12	+0.51	0.65	+3.12	1.05	628
8	loops and decoupler only	2.14	+29.46	−14.71	0.48	1522	22.79	+43.06	0.89	−16.59	4.30	1496
9	row 2, slip-aware limit	1.50	−0.91	−0.05	0.26	1633	6.29	−1.36	0.00	−0.77	0.52	2244
10	MPC, slip-aware limit	2.11	+0.82	+1.11	0.52	568	4.85	+0.76	0.40	+1.65	0.74	951

A ratio above 1 slipped the wafer. No lamp reaches its rating on case 3's recipe, where the slip-aware limit never acts. Each MPC row solved 528 programs on case 3's wafer and 511 on case 9's (516 in row 10), none was rejected, and a solve took 2.1 to 2.4 ms on average in Node on a shared host.

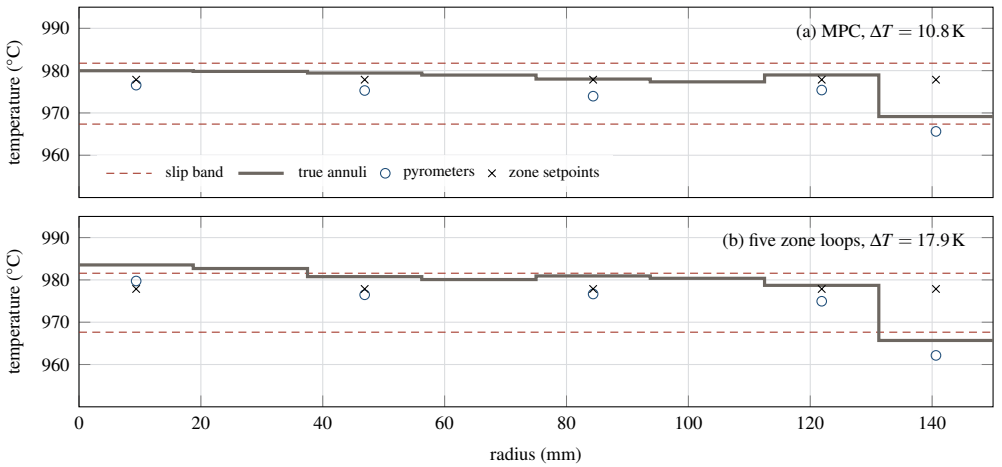


Figure 9.6 Radial profile of the first wafer of case 9 at $t = 24.2$ s, as the 400 °C/s lead trajectory reaches the peak and the setpoints stand at 977.9 °C, under (a) the MPC and (b) the five zone loops with the feedforward and the decoupler. Steps are the true annulus temperatures, circles the pyrometers, crosses the zone setpoints, and the dashed lines the slip band, as wide as the limit at the hottest annulus and centered on the profile. The MPC holds its edge lamps at their rating and the inner zones at 58 to 73 %, for 10.8 K against a limit of 14.4 K. The loops have cut every zone to 26 to 57 % with the center at 983.5 °C and the edge annulus at 965.7 °C, 17.9 K against 13.9 K.

TRY IT IN THE APP **9 · MPC Against the Zone Loops on a Fast Ramp**

Press Run and watch the Radial view during the ramp at about 23 s. Under the MPC the radial difference peaks at 10.89 K and the wafer does not slip, but its sheet resistance is 3.34 % high. Press Reset, set Zone control to Five zone loops and run again: the difference reaches 17.27 K and the wafer slips.

The counterfactual rows say where the smaller spread comes from. It survives the removal of each mechanism that the MPC adds to the loops, at 11.12 K without the weight on neighboring spots, 11.69 K without the lamp limits in the plan and 6.76 K with the preview off, so it comes from

Table 9.8 Graded Lot Under the Zone Loops and the MPC, at the Shipped Seed

Configuration	In spec	Worst R_s (%)	Worst peak (K)	Slips
as shipped	0	15.34	27.83	23
suggested loops	25	1.23	2.32	0
MPC, constant ε	9	28.45	30.94	0
MPC, film-stack table	19	2.67	1.14	0
retuned loops, table	15	3.16	1.88	0

planning the five lamps together on the model while they saturate. The spread weight decides only the verdict on slip. Graded from 0 through 0.25, 0.5 and 1 to 2/K^2, it moves the ratio from 1.05 (a slip) through 0.92, 0.83 and 0.79 to 0.77 while the spread stays between 10.6 and 11.1 K. Without the preview the plan, which then sees only the setpoint of the moment, lags the ramp and keeps the wafer more uniform at the same budget (+3.35 %). The shipped structure already receives the same recipe, through its feedforward, in row 2.

The supervisor's own answer is better for the product. With the slip-aware limit on, the decoupled loops stretch the ramp from 1.13 to 1.94 s, hold the difference to 6.29 K (0.52 of the limit) and make a wafer 0.77 % low, inside the band, which beats every MPC row. A slower ramp asks less of the edge lamps, and no plan of the same fast ramp can do that. With the same limit the MPC has the smallest spread of all, 4.85 K, but is still 1.65 % high.

On the nominal recipe the MPC loses. Its spread is 2.11 K against 1.50 K, its true center peak 0.82 K over the recipe instead of 0.91 K under, and its sheet resistance 1.11 % high against −0.05 %, with only its lamp travel lower (568 against 1633 kW). The budget is lost at the end of the soak. Of the 0.0508 s of equivalent budget that the MPC's wafer lacks (3.8 % of the loops'), 0.0319 is lost in the first 0.5 s of the cool and 0.0169 in the next 1.5 s, against 0.0016 in the soak. After the cut the setpoint follows the model's free cooling through the lamp lag, which a wafer with its lamps off cannot follow exactly, and the least-squares plan, which knows the lamps cannot go below off, spreads that error across the corner. It lowers the lamps from 46.1 to 25.8 and 12.8 kW over the last 0.2 s of the soak, where the loops hold 44 kW, and the true center then falls 5.2 K below its setpoint (the loops' 0.5 K). Allowing negative lamp power after the cut, a counterfactual that makes the setpoint reachable in the model and changes nothing else, brings the sheet resistance to +0.41 % on case 3 and +2.49 % on case 9, where the spread stays at 10.93 K. Removing the weight after the cut gives +0.31 %, while a move weight ten times lighter, a hundred times lighter or ten times heavier gives +1.14, +1.14 and +1.18 %, and one move block per step +1.12 %.

The emissivity error stays the lesson under every structure, because the MPC reads the same pyrometers. On case 4's nitride wafers it makes +27.18 % on the first two wafers against +26.31 and +26.29 % for the decoupled loops and +26.18 % for the retuned loops, with its true center 29.9 K below the peak instead of 33.0, and its trend looks just as perfect. On the graded lot, Table 9.8, the MPC with the film-stack table puts 19 wafers in specification and fails, because wafers 10 to 15 on the nitride stack end 2.3 to 2.7 % high, although no wafer slips and every true center peak is within 1.14 K. The suggested loops put all 25 in specification.

Every existing case was also run for 1501 s under each structure, to find the worst case of the hottest annulus, the slip ratio and the sheet resistance. No plan was rejected anywhere. The MPC never slips a wafer that the loops keep, and it holds every wafer they slip except the weak edge of case 6, which the supervisor holds after the first wafer under both. It is nonetheless less safe than the best shipped structure in four cases (4, 5, 8 and the suggested lot), where its hottest annulus is 0.5 to 3.3 K higher, late in the soak, and its worst slip ratio higher (0.54 against 0.19, 0.65 against 0.26, 0.81 against 0.27 and 0.67 against 0.54), although below 1 wherever the loops' is. In case 2 its slip ratio of 0.65 also exceeds the five zone loops' 0.48, with its hottest annulus 28 K lower. Its worst wafers are also further from the target, 5.16 % against 2.87 % in case 8 and 2.74 % against 1.43 % in case 5.

The application states this in the texts of case 9 and in its model description.

The cost fits the browser. On the page's own frame loop in case 9, timed twice on a shared host (load averages of 3 to 6 and of 12 to 13), a decision takes 2.43 to 2.57 ms on average and at most 11.3 to 13.5 ms, and at real time the solves use 4.9 to 5.1 % of a processor core and all the stepping 8.9 to 9.5 %, at 60 frames per second. At 20 times real time a frame guard holds the run to 16.6 to 17.0 times at about 42 frames per second, where the zone loops run the full 20 times on 4.6 to 5.2 % of a core, so a lot of 1500 s takes 75 s, and the results are identical at 1, 5 and 20 times real time.

9.7 Technology in Practice

The loops of Section 9.5 run on five pyrometers under the wafer, a lamp head of five zones, the tool's own controller and a probe after the chamber. Figure 9.7 draws one zone with that hardware.

What Measures It

Each zone reads one pyrometer, which the model of Section 9.3 reduces to a band-averaged reading behind a 5 ms lag with 0.3 K of colored noise over 350 to 1250 °C, converted with the emissivity it is told. Its radiation is gray and band averaged and its reflector adds no cavity enhancement, so its emissivity error is about three times that of a narrow-band instrument.

> **TECHNOLOGY Sensor: A Lightpipe Radiation Thermometer**
>
> A straight rod of quartz or sapphire looks at the back of the wafer and carries its glow through an optical fiber to a photodetector. Most such thermometers read a band from 0.9 to 1 µm, where silicon is opaque, so lamp light passing through the wafer stays out [13]. Against a blackbody one is calibrated with an uncertainty below 0.3 °C. In a tool the wafer's emissivity and its reflections off the chamber must be accounted for, or the uncertainty grows by orders of magnitude, and a calibration in the tool reached 2.1 °C or less [18]. The cold rod also cools the spot it reads: depressions of up to 25 °C were seen with its tip 2 mm from the wafer [22]. A wrong emissivity leaves no mark on the trend.

The model's pyrometers never cool their spots and err only through the emissivity they are given and the window's radiance, the faults of cases 4 and 8. Its probe reads every wafer at once with a repeatability of 0.3 %, where a fab measures later and on sampled wafers.

What Moves It

Each zone is modeled as a power proportional to its command behind a filament lag of 0.20 s, with ratings from 9 to 52 kW, 145 kW in all, and an idle filament holds 1 % of each rating.

> **TECHNOLOGY Final Element: Tungsten-Halogen Lamps on Thyristor Power**
>
> A tungsten-halogen lamp is a tungsten filament in a quartz tube filled with halogen gas. With the quartz above 250 °C, the halogen cycle keeps tungsten off the envelope, which prevents blackening and lengthens the lamp's life. The filament runs near 2900 K, and the quartz, warmed to about 400 °C, radiates heat of its own. A university chamber of 2005 carried 37 such lamps in three concentric zones, and its controller sent analog commands to the lamp power supplies [17]. Alternating current to a lamp can be throttled by phase control: the later in each half-cycle a thyristor is fired, the less average power the load receives, and large power circuits favor silicon-controlled rectifiers [23]. A zone that has lost lamps shows up as a command held at its rating.

The model's proportional power stands for the average such a controller delivers, and its aged zones stand for lost lamps.

Where the Controller Runs

In a university chamber of 2005 a programmable logic controller read the thermocouples and pyrometers, ran the PID loops and sent the lamp commands, while a workstation above it ran

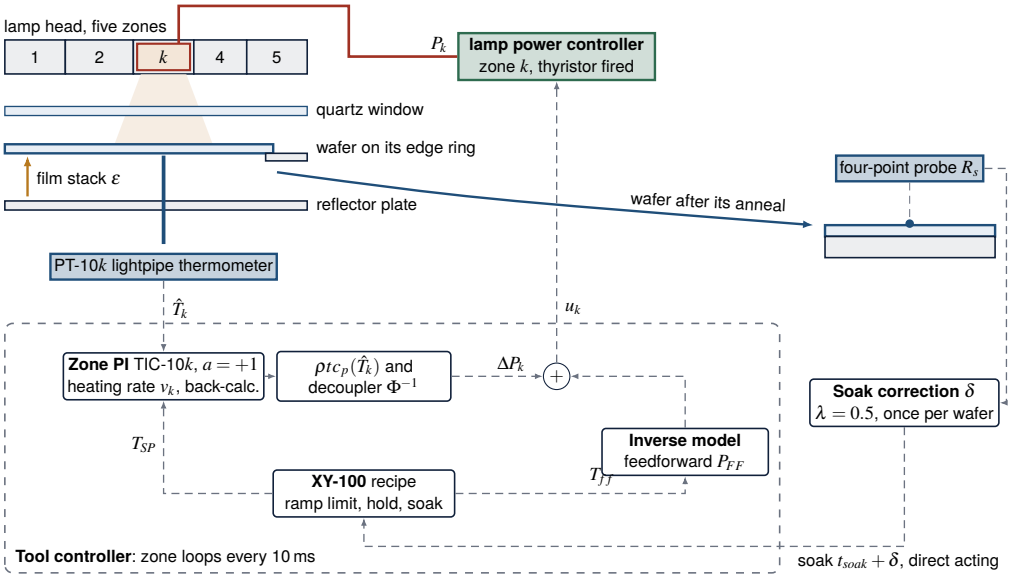


Figure 9.7 One of the five zone loops of the simulated chamber drawn with its hardware, and the soak correction beside it. The lightpipe thermometer PT-10k under the wafer feeds TIC-10k, a velocity-form PI loop with $a = +1$ every 10 ms whose output is a heating rate. The areal heat capacity and the decoupler turn it into lamp power, the inverse-model feedforward from the recipe of XY-100 is added, and the sum drives the power controller of lamp zone k. Once per wafer the probe's sheet resistance lengthens or shortens the soak, a direct-acting correction. Notice that no instrument of the inner loop sees the true wafer temperature, and that a film stack changes what the thermometer reads.

identification and supervisory tasks. Its multiloop PI tests ran at 1 and 10 Hz, and the faster rate gave less overshoot [17]. Control loops in operational technology run with cycle times from milliseconds to minutes [24]. The app integrates every 1 ms, runs its zone loops every 10 ms and the MPC every 50 ms, and corrects the soak once per wafer, every 60 s.

> **IN PRACTICE** **Calibrating the Thermometers Against a Thermocouple Wafer**
>
> Thermometers in a lamp tool are calibrated in place with a wafer that carries thermocouples. The difference between a thermometer's display and its thermocouple then corrects the later readings, and the calibration should use the same spacing between wafer and reflector as production [18]. A film on the measured side of a wafer changes its emissivity. Such wafers are screened out by hand, and a 2005 study asked for a systematic alarm on the tool instead [17].

In such a controller the structure is five PI blocks with back-calculation, a five-by-five matrix block for the decoupler, a feedforward block that reads the recipe, the sequence of XY-100 with its ramp limit and hold, and a run-to-run update on the soak.

What Is Modernizing It

The first development is multivariable, model-based control. By 2003 controllers built on physics-based models reduced to low order ran on commercial chambers [6], like the MPC of Section 9.6.9, while two years later many tools still kept one loop on a center pyrometer, the architecture of case 2 [17]. The second is shorter annealing. Flash lamp annealing, a non-equilibrium method on the sub-second time scale, has been used in microelectronics mostly after ion implantation, to activate dopants, to recrystallize amorphous layers and to anneal out defects [25]. Measuring the wafer during a flash is the open challenge of Section 9.8.

9.8 Limitations

The thermal model is a teaching model, with eight annuli that leave the rim's own dip unresolved, gray band-averaged radiation that makes the pyrometer error about three times a narrow-band instrument's at 0.95 µm, a fixed lamp absorptivity, a lumped ring and window, and an axisymmetric chamber with no rotation, bow, gas flow pattern or pattern-dependent absorption [5].

The product and the constraint are illustrative. The budget, the sheet resistance law and the slip limit rank thermal histories and are not models of activation, diffusion, junction depth or dislocation. Metrology is immediate and complete, where a fab measures later and on sampled wafers, so the soak correction of Equation (9.14) converges faster than a production loop would. Every size, rating and limit is a teaching value and not a tool's specification.

The MPC's model leaves out radial conduction, the edge ring, the window, the pyrometers' lag and aged lamps, so its gains are 15 to 26 % high at 600 °C and the disturbance estimate carries the rest. Its slope and heat capacity are taken at the reference rather than at the estimate, and the slope comes from the emissivity the pyrometers are told, so it inherits the emissivity error of case 4 unchanged. Its objective weighs the profile against the setpoint and not the thermal budget, which is why a plan that respects the lamps' floor loses budget at every cut, and an objective written on the budget itself would need the Arrhenius term that the plan does not carry.

> **OPEN CHALLENGE** **Measuring a Wafer That Is Hot for Milliseconds**
>
> Annealing has moved below the spike, to flash lamps and lasers that heat and cool a wafer within milliseconds. Real-time temperature control within a wafer and from wafer to wafer is one of the key challenges such tools must meet in production [26]. One prototype hid the flash behind a layer of flowing water under the lamps, read the wafer at 1.95 µm with a thermometer that compensates emissivity from two polarizations, and found its peak reading tracking the sheet resistance whatever the surface emissivity [27]. Progress would be such a reading on every product wafer, an independent sensor.

One limit is a property of the application, not of the model. The choice between a lot and one held wafer belongs to each case and is shown read-only, because a held wafer and a lot start from different chamber states, so a student who wants the held wafer loads case 1. The lot size applies at once, never below the wafers already annealed.

9.9 Conclusion

This chapter studied a spike anneal chamber, a browser plant whose loops read pyrometers while its product is computed from the true temperatures. The open-loop time constant falls from 13.7 to 2.4 s between 400 and 1050 °C, yet loops that ask for a heating rate answer a 10 K step in 0.33 s at all four temperatures tried, so the cubic schedule is offered and not needed. Feedback alone lags the 250 °C/s ramp by 65.2 K and overshoots by 29.9 K, and the inverse-model feedforward brings the lag to 2.9 K. Five independent loops pull the radial difference to 34.8 K, which the decoupler holds to 2.1 K. A 10 % emissivity error leaves the wafer 31.4 K below its pyrometers and 26.3 % high in sheet resistance until a table or a calibration wafer corrects it, and the graded lot improves from none of 25 wafers in specification to all 25. A five-zone MPC holds the 400 °C/s ramp of case 9 to 10.9 K without slip where the decoupled loops reach 17.3 K and slip, but it leaves the wafer 3.3 % high, and the supervisor's slower ramp makes a better wafer than either.

The model overstates the pyrometer error and uses illustrative product laws. A fuller model would add wafer rotation, spectral emissivity with a narrow-band pyrometer and a sampled metrology queue. The chamber is an accessible plant for learning that a loop on its setpoint has measured its sensor, not its product.

KEY POINTS

- A loop that holds its pyrometer on the recipe has measured its sensor, not the wafer it anneals.
- Feedback alone lags the 250 °C/s ramp by 65.2 K and overshoots by 29.9 K, and the inverse-model feedforward brings the lag to 2.9 K.
- A 10 % emissivity error leaves the wafer 31.4 K below its pyrometers and 26.3 % high in sheet resistance.
- A five-zone MPC holds the 400 °C/s ramp without slip where the decoupled loops slip, but the supervisor's slower ramp makes a better wafer than either.

Where else this applies. A pyrometer that sees only the surface in Chapter 2, coupled heating zones in Chapter 5, and a product temperature no instrument reaches in Volume 5, Chapter 11.

References

[1] R. Singh, "Rapid isothermal processing," *Journal of Applied Physics*, vol. 63, no. 8, pp. R59–R114, 1988.

[2] F. Roozeboom and N. Parekh, "Rapid thermal processing systems: A review with emphasis on temperature control," *Journal of Vacuum Science & Technology B: Microelectronics Processing and Phenomena*, vol. 8, no. 6, pp. 1249–1259, 1990.

[3] A. Agarwal, A. T. Fiory, H.-J. L. Gossmann, C. S. Rafferty, and P. Frisella, "Ultra-shallow junction formation by spike annealing in a lamp-based or hot-walled rapid thermal annealing system: effect of ramp-up rate," *Materials Science in Semiconductor Processing*, vol. 1, no. 3–4, pp. 237–241, 1998.

[4] P. J. Timans, "Emissivity of silicon at elevated temperatures," *Journal of Applied Physics*, vol. 74, no. 10, pp. 6353–6364, 1993.

[5] J. P. Hebb and K. F. Jensen, "The effect of multilayer patterns on temperature uniformity during rapid thermal processing," *Journal of The Electrochemical Society*, vol. 143, no. 3, pp. 1142–1151, 1996.

[6] A. Emami-Naeini, J. L. Ebert, D. de Roover, R. L. Kosut, M. Dettori, L. M. L. Porter, and S. Ghosal, "Modeling and control of distributed thermal systems," *IEEE Transactions on Control Systems Technology*, vol. 11, no. 5, pp. 668–683, 2003.

[7] C. D. Schaper, Y. M. Cho, and T. Kailath, "Low-order modeling and dynamic characterization of rapid thermal processing," *Applied Physics A Solids and Surfaces*, vol. 54, no. 4, pp. 317–326, 1992.

[8] Y. M. Cho and P. Gyugyi, "Control of rapid thermal processing: a system theoretic approach," *IEEE Transactions on Control Systems Technology*, vol. 5, no. 6, pp. 644–653, 1997.

[9] K. S. Balakrishnan and T. F. Edgar, "Model-based control in rapid thermal processing," *Thin Solid Films*, vol. 365, no. 2, pp. 322–333, 2000.

[10] T. F. Edgar, S. W. Butler, W. J. Campbell, C. Pfeiffer, C. Bode, S. B. Hwang, K. S. Balakrishnan, and J. Hahn, "Automatic control in microelectronics manufacturing: Practices, challenges, and possibilities," *Automatica*, vol. 36, no. 11, pp. 1567–1603, 2000.

[11] E. Sachs, A. Hu, and A. Ingolfsson, "Run by run process control: combining SPC and feedback control," *IEEE Transactions on Semiconductor Manufacturing*, vol. 8, no. 1, pp. 26–43, 1995.

[12] D. E. Seborg, T. F. Edgar, D. A. Mellichamp, and F. J. Doyle, III, *Process Dynamics and Control*, 4th ed. Hoboken, NJ: John Wiley & Sons, 2016.

[13] Y. Qu, "Silicon wafer surface temperature measurement using light-pipe radiation thermometers in rapid thermal processing systems," Ph.D. dissertation, The University of Texas at Austin, Austin, TX, 2006. [Online]. Available: https://hdl.handle.net/2152/2794

[14] G. G. Bentini, L. Correra, and C. Donolato, "Defects introduced in silicon wafers during rapid isothermal annealing: Thermoelastic and thermoplastic effects," *Journal of Applied Physics*, vol. 56, no. 10, pp. 2922–2929, 1984.

[15] H. A. Lord, "Thermal and stress analysis of semiconductor wafers in a rapid thermal processing oven," *IEEE Transactions on Semiconductor Manufacturing*, vol. 1, no. 3, pp. 105–114, 1988.

[16] T. Breedijk, T. F. Edgar, and I. Trachtenberg, "A model predictive controller for multivariable temperature control in rapid thermal processing," in *Proceedings of the 1993 American Control Conference*. IEEE, 1993, pp. 2980–2984.

[17] W. Cho, "Temperature control and modeling of rapid thermal processing," Ph.D. dissertation, The University of Texas at Austin, Austin, TX, 2005. [Online]. Available: https://hdl.handle.net/2152/1527

[18] C. W. Meyer, D. P. DeWitt, K. G. Kreider, F. J. Lovas, and B. K. Tsai, "ITS-90 calibration of radiation thermometers for RTP using wire/thin-film thermocouples on a wafer," *AIP Conference Proceedings*, vol. 550, pp. 254–258, 2001.

[19] S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Control Engineering Practice*, vol. 11, no. 7, pp. 733–764, 2003.

[20] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, "Constrained model predictive control: Stability and optimality," *Automatica*, vol. 36, no. 6, pp. 789–814, 2000.

[21] G. Pannocchia and J. B. Rawlings, "Disturbance models for offset-free model-predictive control," *AIChE Journal*, vol. 49, no. 2, pp. 426–437, 2003.

[22] K. G. Kreider, D. H. Chen, D. P. DeWitt, W. A. Kimes, and B. K. Tsai, "Lightpipe proximity effects on Si wafer temperature in rapid thermal processing tools," in *Proc. 11th IEEE Int. Conf. Advanced Thermal Processing of Semiconductors*. IEEE, 2003, pp. 125–129.

[23] T. R. Kuphaldt, *Lessons In Electric Circuits, Volume III: Semiconductors*. Open textbook (CC BY), ibiblio.org, 2023, chapter 7, Thyristors.

[24] K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, A. Hahn, S. Saravia, A. Sherule, and M. Thompson, "Guide to operational technology (OT) security," National Institute of Standards and Technology, Gaithersburg, MD, Special Publication 800-82r3, 2023.

[25] L. Rebohle, S. Prucnal, Y. Berencén, V. Begeza, and S. Zhou, "A snapshot review on flash lamp annealing of semiconductor materials," *MRS Advances*, vol. 7, no. 36, pp. 1301–1309, 2022.

[26] S. Chen, J. Hebb, A. Jain, S. Shetty, and Y. Wang, "Wafer temperature measurement and control during laser spike annealing," in *Proc. 15th Int. Conf. Advanced Thermal Processing of Semiconductors*. IEEE, 2007, pp. 239–244.

[27] Y. Yamada, T. Aoyama, H. Chino, K. Hiraka, J. Ishii, S. Kadoya, S. Kato, H. Kiyama, H. Kondo, T. Kuroiwa, K. Matsuo, T. Owada, T. Shimizu, and T. Yokomori, "In situ Si wafer surface temperature measurement during flash lamp annealing," *Japanese Journal of Applied Physics*, vol. 49, no. 4S, p. 04DA20, 2010.

9.10 Exercises

9.1 ⌕ Concept check · 2 min

The open-loop time constant of the wafer falls from 13.7 s at 400 °C to 2.4 s at 1050 °C. Yet a 10 K setpoint step on the held wafer of case 1 in automatic reaches 63 % in 0.33 s at all four temperatures tried. Why does the response not change with temperature?

(a) The pyrometer lag of 5 ms sets the response at every temperature.
(b) The loops ask for a heating rate, and their gain dwarfs the wafer's radiative rate.
(c) The cubic gain schedule is on in case 1 and cancels the change of the conductance.
(d) The heat capacity rises with temperature about as fast as the conductance does.

9.2 ⌕ Concept check · 1 min

In case 4 the pyrometers assume an emissivity of 0.70 on nitride-capped wafers whose emissivity is 0.77. On a real tool, which signal reveals the error?

(a) The center pyrometer's trend, which runs above its setpoint through the soak.
(b) The difference between the pyrometers, which grows on the ramp.
(c) The sheet resistance measured on the wafer after it leaves the chamber.
(d) The lamp commands, which reach their ratings during the ramp.

9.3 ⌕ Concept check · 2 min

In case 3 the static decoupler alone cuts the largest radial difference on the ramp from 34.8 K to 2.1 K, yet the first wafer's sheet resistance comes out 14.7 % low. What explains the low sheet resistance?

(a) The decoupler drives zone 5 to its rating, which leaves the rim cold through the soak.
(b) The decoupler inverts the gain matrix of 1050 °C only, so it is wrong on the ramp.
(c) The decoupler slows every loop, so the wafer never reaches the recipe peak.
(d) The wafer still lags the ramp by about 65 K as one piece and then overshoots.

9.4 Theory and calculation · 6 min

The pyrometers convert radiance with an assumed emissivity ε_{set}, and in the band-averaged model a loop that holds the reading on T_{SP} holds the wafer at $T = T_{SP}(\varepsilon_{set}/\varepsilon)^{1/4}$ in kelvin (Equation (1.1)). The setpoint is 1050 °C and $\varepsilon_{set} = 0.70$. Give temperatures to 0.1 K.

(a) Find the true wafer temperature for a nitride cap with $\varepsilon = 0.77$, and how far it lies below the setpoint. *Checkpoint:* 1018.8 °C, 31.2 K low

(b) Find the true wafer temperature for an oxide stack with $\varepsilon = 0.63$. *Checkpoint:* 1085.3 °C

(c) A narrow-band pyrometer at $\lambda = 0.95\,\mu\text{m}$ follows Wien's law, $\Delta T/T = (\lambda T/c_2)\,\Delta\varepsilon/\varepsilon$ with $c_2 = 14\,388\,\mu\text{m K}$. Find its relative and absolute error at 1050 °C for $\Delta\varepsilon/\varepsilon = 0.10$, and the ratio of that error to the one of part (a), to two decimals. *Checkpoint:* 0.87 %, 11.6 K, ratio 0.37

9.5 Theory and calculation · 6 min

The recipe ramps the wafer from 600 to 1050 °C at 250 °C/s. Near the peak the wafer's heat capacity is 126.0 J/K and its conductance K is 53.1 W/K. Each zone loop outputs a heating rate $v = K_c e$ plus its integral, with $K_c = 4\,\text{K/s}$ per kelvin.

(a) Find the duration of the ramp and the absorbed power the ramp alone asks for near the peak. *Checkpoint:* 1.8 s, 31.5 kW

(b) Ignore the integral action, the losses and the lamp lag. What error must a proportional loop carry to ask for 250 °C/s? Compare it with the center lag of 65.2 K that case 2 measures without feedforward, and say why the integral cannot remove the lag on this ramp ($\tau_I = 1.5\,\text{s}$). *Checkpoint:* 62.5 K

(c) Find the open-loop time constant $m_w c_p/K$ at 1050 °C, to two decimals, and compare it with the ramp duration. *Checkpoint:* 2.37 s

9.6 Theory and calculation · 8 min

The budget of Equation (9.7) has $E_a = 3.5\,\text{eV}$, $k_B = 8.617 \times 10^{-5}\,\text{eV/K}$, $m = 0.30$ and $\theta_T = 1.3365\,\text{s}$. The soak correction of Equation (9.14) uses $\lambda = 0.5$.

(a) At $T = 1323.15\,\text{K}$, find the relative change of the budget rate per kelvin, $E_a/(k_B T^2)$, the resulting change of R_s per kelvin of peak, both in percent to two decimals, and the peak error that matches the $\pm 2\,\%$ band, to 0.1 K. *Checkpoint:* 2.32 %, 0.70 %, 2.9 K

(b) The probe reads a wafer 2.87 % high. By how much does the correction lengthen the next soak, in seconds to three decimals? Is the correction direct or reverse acting in the convention of Section 1.6.2? *Checkpoint:* 0.063 s, direct acting

(c) One more second of soak changes $\ln R_s$ by $-m/\theta_T$, so the correction is a loop $e_{k+1} = e_k - K e_{k-d}$ on $e = \ln(R_s/R_{s,T})$, with d the wafers by which the probe result is late (Chapter 1's metrology delay, Equation (1.6)). Find K. Using Equation (1.7), is the correction stable if the result arrives two wafers late, and if it arrives three wafers late? *Checkpoint:* $K = 0.5$; stable at two, unstable at three

9.7 App challenge · 8 min

Load scenario 2, "One loop on the center", default seed, at speed 5. Run to $t = 61\,\text{s}$, the end of the first wafer.

(a) Read the first wafer's true mean sheet resistance on the R_s last badge. Export the CSV and find the largest value of `sp1_C` minus `pyro1_C` over the rows whose `phase` is `ramp`. Give the sheet resistance to 0.1 Ω per square and the lag within 0.5 K. *Checkpoint:* 355.3 Ω/sq, 64.9 K

(b) Press Reset, set Model feedforward On on the TY-100 card before pressing Run, run to 61 s again and read the same two values, the lag within 0.3 K. *Checkpoint:* 404.7 Ω/sq, 2.6 K

(c) The wafer of part (a) lagged its setpoint on the ramp, yet its sheet resistance is low. Explain why, from the trend of the center pyrometer at the top of the ramp. *Checkpoint:* an overshoot of about 30 K over-anneals the wafer

9.8 App challenge · 8 min

Load scenario 4, "Emissivity wrong by 10 %", default seed, at speed 5. Run to $t = 61$ s and export the CSV.

(a) For the rows with `wafer` equal to 1, find the largest `T_ann0_C` (the true center annulus) within 0.1 K and the largest `pyro1_C` within 0.3 K. Read the sheet resistance on the R_s last badge to 0.1 Ω per square. *Checkpoint:* 1016.9 °C, 1048.2 °C, 505.2 Ω/sq

(b) Apply $T = T_{read}(0.70/0.77)^{1/4}$ in kelvin to the largest pyrometer reading of part (a) and compare the result with the largest `T_ann0_C`. Then estimate the sheet resistance error from 0.70 % per kelvin of peak and a peak 31.2 K low, and say why the estimate falls short of the measured error. *Checkpoint:* 1017.1 °C; about 22 % against 26.3 %

(c) Press Reset, set Pyrometer emissivity to Film-stack table before pressing Run, run to 61 s and read the badge again. *Checkpoint:* 402.3 Ω/sq

9.9 App challenge · 10 min

Load the challenge, "A lot of 25 wafers", default seed, at speed 20 (the 1500 s lot takes about 75 s). The run stops by itself when the lot is complete.

(a) Run the lot as shipped. Read the In spec, Worst R_s error, Worst peak error and Slip flags tiles. *Checkpoint:* 0, 15.34 %, 27.8 K, 23

(b) Press Reset. Before Run set Pyrometer emissivity to Film-stack table, Model feedforward On and, in Advanced, Static decoupler On. Run the lot and read the same four tiles and the Lamp energy tile. *Checkpoint:* 25, 1.23 %, 2.3 K, 0, 3.4 kW h

(c) Which of the three changes carries wafers 10 to 15, and why can the other two not replace it? Use Table 9.6. *Checkpoint:* the film-stack table

9.10 Application area · 3 min

In case 6, zone 5 delivers 35 % of its output and the supervisor holds the lot after the first wafer. A shift lead proposes to keep the tool running by slowing the recipe to 50 °C/s with the slip-aware limit and the hold off. Using the chapter's measurements, what do you advise, and why is holding the lot the right control action?

9.11 Application area · 3 min

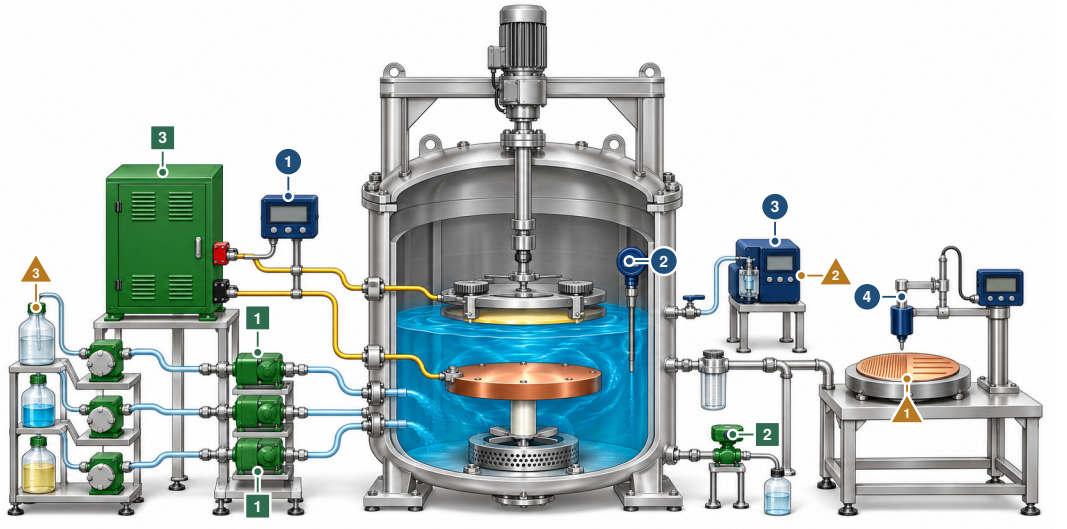
A fab introduces a new nitride-capped product. Compare the film-stack table and a calibration wafer as sources of the pyrometers' emissivity: what does each cost, what does each rely on, and when does each go out of date?

9.12 Review the agent · 5 min

An AI assistant reviews case 9 and writes: "The MPC holds the 400 °C/s ramp to a radial difference of 10.9 K, 0.79 of the slip limit, where the decoupled zone loops slip at 1.77 times it. The MPC is therefore the better controller for this chamber. Switch the challenge lot to Zone control MPC with the film-stack table and every wafer will be in specification." Find the flaw and say how you would check the claim in the app.

CHAPTER 10

Copper Plating Bath Simulator Dosed by Amp-Hours and Analyzed Every Two Hours



The controller moves 1 three pumps that dose the accelerator, suppressor and leveler
2 a bleed that drains bath to waste, refilled with fresh solution 3 the plating time of each wafer, run by the power supply

The controller reads 1 a counter of the charge the power supply delivers 2 a thermometer in the bath
3 an analyzer that tests a bath sample every two hours 4 a probe that measures each plated wafer's thickness

What disturbs it 1 lots of thick-copper wafers that take twice the charge 2 the analyzer out of service
3 a new lot of additive concentrate that the bath uses faster

Notice that the pumps can follow the charge counter at once, while the analyzer at the right answers every two hours about a sample more than forty minutes old. Illustration, not to scale.

IN PLAIN TERMS **Feeding a Copper Bath by Its Charge**

The main idea. Every copper wire in a chip is grown in a bath of copper and acid, one silicon disk at a time. Three trace chemicals make the copper fill each hole from the bottom up, and the electric charge that plates each disk uses them up.

The challenge. The bath is tested every two hours, and each test shows it as it was over forty minutes before. Worn-out chemicals also build up and shrink the mix in which holes fill, and only fresh solution can dilute them.

Why it matters. Dosing by the charge meets a change of product at once, which a slow test cannot.

THREE WAYS TO READ THIS CHAPTER

Explore	12 min	The first page, the history boxes of Section 10.2, the fact boxes of the Introduction and of Section 10.7, the try-it boxes of Section 10.7, and the key points
Study	50 min	The whole chapter with its pause points and the twelve exercises
Practice	15 min	The first page, Section 10.7, Section 10.8, Section 10.9 with its open challenge

The control question A copper plating bath consumes its additives with every amp-hour, while its only analysis arrives every two hours and describes the bath of 42 minutes before. How much of the dosing can such an analysis carry, and what must the amp-hour counter and the bleed do instead?

Learning objectives

- Compute a wafer's charge, bulk step and thickness from Faraday's law, and the additive dose that holds the bath at a given load.
- Explain why a feedforward on the amp-hours answers a lot change at once while a trim on the analysis first sees it hours later.
- Use the report-to-report model to explain why the sampling and the analysis age, not the chemistry, set the gain limit of the trim.
- Price a setpoint of the breakdown products in virgin make-up with the make-up balance, and decide between bleed-and-feed and a dump.
- Judge when an analysis should be rejected against the controller's own model of the bath.

Prerequisites Sections 1.4.3 (run-to-run correction), 1.4.4 (late measurement), 1.4.7 (feedforward), 1.6.2 (velocity-form PI) and 1.6.4 (scenarios). Chapter 6 is a useful first look at a correction made on late results.

The app https://apmonitor.com/apps/copper-ecd-bath/

See also Volume 5, Chapter 3 and Volume 5, Chapter 4 for loops behind slow or late analyzers. Volume 3, Chapter 6 for a dose computed from the load with a slow quality trim. Volume 8, Chapter 9 for corrections made while earlier results are still in flight.

10.1 Introduction

Every copper wire and via of a modern chip is plated. Full copper wiring in a CMOS technology was reported in 1997 [1], and in the damascene process that followed each metal level is etched into the dielectric, lined with a seed layer and filled by electroplating from an acid copper sulfate bath [2]. The features fill without a void only because copper grows faster at their bottoms than at their walls and mouths, and that bottom-up growth is the work of organic additives. A suppressor adsorbs with chloride and inhibits the deposit [3], an accelerator displaces it and gathers at the bottom of a filling feature as the feature's area shrinks [4], and a leveler competes with the accelerator where the current crowds [5]. The filling of trenches and vias in the presence of additives has its own theory [6], the accelerator decomposes in these chemistries [7], and the reliability of finished copper interconnects has been reviewed at length [8].

> ✦ DID YOU KNOW? **Six Levels of Copper in 1997**
>
> The first fully integrated copper wiring of a chip technology, reported in 1997, stacked up to six levels of copper lines. The finest were laid at a pitch of 0.63 µm, above transistors of a technology below 0.25 µm [1].

What the simulator sets out to correct is the belief that a bath whose analyses sit on their setpoints will fill its vias. The additives are analyzed by cyclic voltammetric stripping (CVS) [9], and the application's analyzer takes its sample once every 2 h from a line that holds 2 min of bath and publishes the result 40 min after the draw, so every result describes the bath as it stood at least 42 min before it arrived. In the lot change of case 4, dosed on feedback alone, the result reported at 4:40 still reads 5.96 mL/L, because its sample left the bath just before the thick-copper wafers started at 4:00. The first result to show the doubled consumption arrives at 6:40, from a sample drawn when the accelerator was already 6.5 % low, and 97 of the day's 400 wafers void. The analysis also says nothing about the breakdown products, which set the width of the window the accelerator-to-suppressor ratio must stay inside. In case 2 the dosing holds the accelerator at 5.84 mL/L and the suppressor at 2.000 mL/L all day while the breakdown products climb from 27 to 31.4 mL/L and 234 wafers void.

A transfer function from a dosing pump to an analyzer reading carries neither the age of the reading nor the inventory that closes the window.

Pause. Try Exercise 10.1 before reading on.

The class of problem is general. A composition analyzed seldom and late is also the situation of a fermentation whose key concentrations come from delayed offline assays [10], and of any inferential loop built on slow, multi-rate primary measurements [11]. A species that no actuator removes except a purge is the situation of the inerts and byproducts of a plant with recycle [12]. A sparse analysis that a controller will act on for hours has to be screened for gross errors against what the process should have produced [13], and a recipe corrected between runs by an exponentially weighted average (EWMA) of its errors is the run-to-run controller of semiconductor manufacturing [14], the subject of Section 1.4.3 and of Chapter 6. A reagent consumed in proportion to a measured load is dosed forward in ratio to it, which is textbook practice [15]. The bath puts all of these in one vessel.

However, the electrochemistry literature treats the additives as a mechanism of fill [4, 5] and their analysis as a method [9], and the control literature treats sampled analyses, purges and run-to-run correction as separate structures [11, 12, 14], so neither offers a bath on which a result can be watched arriving 42 minutes late while a window it does not report closes. This chapter works out what a sparse and late analysis can and cannot do for a copper plating bath, on a browser simulator in which every wafer is plated by an exact charge, the analyzer and the laboratory report late and can be wrong, and each wafer's true fill is recorded separately from anything an instrument says. Faraday's law, amp-hour feedforward, the PI trim, the EWMA filters, the gross-error gate and bleed-and-feed are all established practice. Eight cases and a graded day separate a load from a model error, a late result from a wrong one, and a dose from an inventory.

Section 10.2 gives the history of the domain. Section 10.3 sets out the bath as modeled: a 39-state model of seven species, the bath temperature and 31 cumulative ledgers that close to 3×10^{-12}, with discrete wafers, a truth log per wafer, and the framework's shared sampled analyzer and bias update, in the detail an independent rebuild needs. Section 10.4 derives the make-up balance, which divides the breakdown products removed by the concentration at which they leave and shows that in a bath whose byproducts accumulate linearly a dump cycle cannot cost less than bleed-and-feed. Section 10.5 describes the controller. Section 10.6 builds a four-state report-to-report model of the sampled trim, whose spectral radius reaches one at a gain of 210 mL/h per mL/L and whose decay per analysis and period agree with the noise-free engine within 0.003 and 0.9 %. Section 10.7 reports the measured cases, Section 10.8 describes the hardware such a bath runs on, and Sections 10.9 and 10.10 state the limitations and conclusions.

10.2 Brighteners, Stripping Analysis and Damascene Fill

The additives were measured electrochemically two decades before their part in filling a feature was modeled, by the stripping method whose beginnings the history box below tells [9, 16]. The simulated analyzer stands for an instrument of that kind, a sample analyzed in tens of minutes on a schedule of hours. The chemistry behind the signal was mapped in the years that followed. Polyethylene glycol was shown to inhibit copper deposition in the presence of chloride ion [3], the pairing that serves as the suppressor of the damascene bath [5].

> **HISTORY** **Measuring a Brightener by Stripping Copper (1978 to 1985)**
>
> In 1978 a voltammetric stripping method was published for the brighteners and levelers of plating baths. The potential of an inert rotating electrode is cycled in the bath, so that copper is plated on it and stripped off again. The charge needed to strip the copper follows the rate of plating, and so the concentration of the additive. In copper pyrophosphate baths the method reached a precision of about 0.1 parts per million over a range from 0 to 2.5 [16]. A pulse version of 1985 found the absolute concentration of an additive in acid copper baths, fresh and in production, that carried contaminants from circuit board plating [9].

Table 10.1 Variables and States of the Bath, With Setpoints and With the Values Case 4 Begins From

Variable	Role	Design value	Range
S_{CVS}, P_{CVS}, L_{CVS}	CVs, AIC-S, P, L	6.0, 2.0, 3.0 mL/L	3–10, 1–4, 1–6
$\hat{B}$	CV, FIC-BLEED	22 mL/L	4–30
h	CV, R2R	1.00 µm	0.6–1.3
Cu_{lab}, H_{lab}	CVs, AIC-CU	40, 10 g/L	30–50, 5–20
m	quality, true	0.267	void below 0
D_S, D_P, D_L	MVs, dosing pumps	57.4, 2.04, 7.20 mL/h	$\leq$ 150, 12, 40
b	MV, bleed	1.76 L/h	0–4
t_b	MV, bulk step	125.2 s	60–146
w, $\dot{m}_{Cu}$, $\dot{m}_H$	MVs, titration	0	—
thick share	DV, lot mix	0 %	0–100
ΔT_{ch}	DV, chiller drift	0 °C	−3 to 12
$\Delta\eta_a$, $\Delta\eta_c$	DVs, efficiencies	0 %	−5–10, −8–2
CVS down, idle	DVs, switches	off	on, off
outlier	DV, S_{CVS} event	—	+3.0 mL/L
overdose	DV, L event	—	+300 mL
S, P, L, B	states	6.0, 2.0, 3.0, 22 mL/L	—
Cl	state	50 mg/L	—
Cu, H, T	states	40, 10 g/L, 25 °C	—
$z_1 \ldots z_{31}$	states, ledgers	0	—

Damascene wiring gave the additives a task with no tolerance for a void, and the theory of how they fill a feature followed within a few years, as the history box after this paragraph tells [1, 2, 4, 6]. The roles of suppressor, accelerator and leveler in damascene plating were reviewed [5]. The decomposition of the sulfonate accelerator in acid copper chemistries was measured directly [7], and it is the origin of the breakdown products this model carries. A leveling agent also governs the cathode quality of copper electrorefining, which Volume 15, Chapter 8 controls at an inferred limit.

> ⧖ HISTORY **Superfilling and Its Theory (the 1990s to 2005)**
>
> Copper electroplating for on-chip wiring was conceived and developed in the early 1990s. The task was to fill trenches and vias without a void or a seam, and plating turned out to fill them far better than a coating of even thickness could. This superconformal growth, called superfilling, depends on the plating additives [2]. In 2000 simulations showed that a poorly controlled additive concentration can leave a larger void than a bath with no additive at all [6]. By 2005 a model in which the local growth rate follows the accelerator's coverage explained the bottom-up fill: the accelerator gathers on the concave bottom of a feature as it advances [4].

The control problem took its present shape with the production bath. A bath that serves a plating tool for a day of wafers loses additive with every amp-hour and gains byproduct with every molecule it loses, while its analysis stays as sparse as the stripping method allows, and the tools of the introduction, from delayed assays to purges, were developed for other processes and carry over. The simulator places them in one bath and adds what no fab records, the true fill margin of every wafer beside the analyses its controllers saw.

10.3 The Bath as Modeled

The simulated tool plates one 300 mm wafer every 216 s, 400 a day, from a 200 L acid copper bath fed by three dosing pumps and by a drum of virgin make-up solution (VMS), a complete fresh bath at the make-up strength of every species. The bath, the plating cell, the bleed, the instruments and every symbol in the equations appear in Figure 10.1.

Table 10.1 gives each variable's role, its value when case 4 starts and its range. Of the 39 states, eight describe the bath, its seven species and its temperature, and 31 are cumulative ledgers of the charge and of what each species has had dosed, consumed, broken down, bled and fed. Integrating the ledgers with the chemistry in one step lets any balance be checked from the state vector alone.

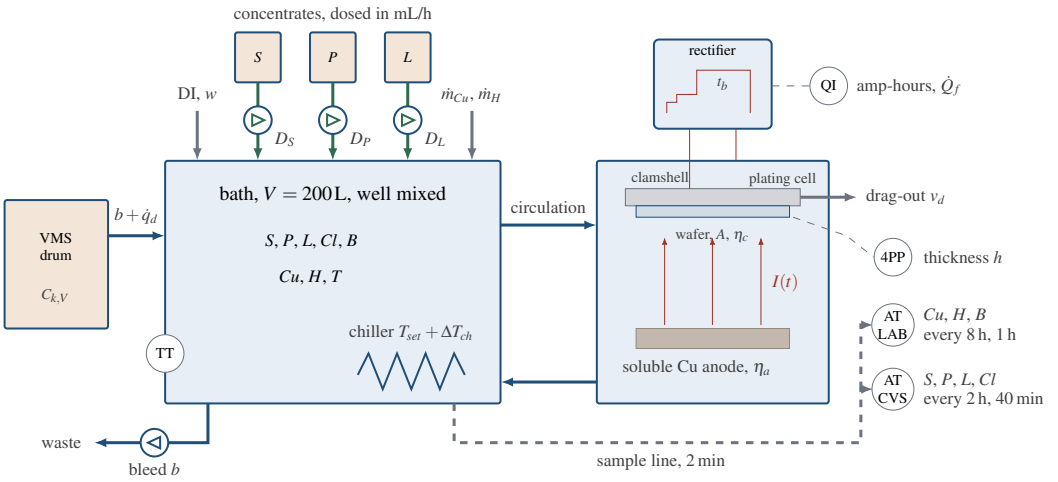


Figure 10.1 The plating bath and the tool it feeds, as modeled. Three pumps dose the accelerator S, the suppressor P and the leveler L as concentrates at D_S, D_P and D_L into a well-mixed bath of volume V, held at $T_{set} + \Delta T_{ch}$ by a chiller and read by TT. The bleed b drains bath to waste, and the VMS drum replaces it and the drag-out v_d of each wafer at the make-up strengths $C_{k,V}$. DI water w and the copper and acid additions $\dot{m}_{Cu}$ and $\dot{m}_H$ are the titration corrections. The plating cell holds the wafer of area A face down in its clamshell over a soluble copper anode, the rectifier drives the three-step current $I(t)$ with a bulk step of length t_b, and the amp-hour counter QI gives the load $\dot{Q}_f$. The CVS analyzer and the laboratory draw from the bath through a 2 min sample line, and the four-point probe 4PP measures each wafer's thickness h. Dashed lines are measurements.

Concentrations are in the units of the tool's panels, mL/L of vendor concentrate for the three additives and their breakdown products, mg/L for chloride and g/L for copper and sulfuric acid, and the engine's time unit is the minute although rates are quoted per hour.

10.3.1 Additives, Chloride and Breakdown Products

Each additive $k \in \{S, P, L\}$ is dosed, consumed at the electrodes in proportion to the charge and to the concentration the surface sees, broken down in the bulk, bled out and fed back with the VMS, which Equation (10.1) writes for the whole bath.

$$\begin{aligned} V\frac{\mathrm{d}C_k}{\mathrm{d}t} &= D_k - c_k I\frac{C_k}{C_{k,V}} - d_k(T)\,V C_k \\ &\quad - (b+w)\,C_k + b\,C_{k,V} \\ d_k(T) &= d_{k,25}\,\mathrm{e}^{E_k(T-25)} \end{aligned} \tag{10.1}$$

The concentration C_k is in mL/L and the dose D_k in mL/h, I is the rectifier current in A, c_k the consumption in mL per A h at the make-up strength $C_{k,V}$, $d_{k,25}$ the breakdown rate at 25 °C in h^{-1} with temperature coefficient E_k in K^{-1} and T in °C, while b is the bleed and w the titration's DI dilution, both in L/h. Chloride is not dosed. It enters only with the VMS and is taken into the deposit at $c_{Cl} = 0.5$ mg per A h, so that $V\,\mathrm{d}Cl/\mathrm{d}t = -c_{Cl}\,I\,Cl/Cl_V - (b+w)\,Cl + b\,Cl_V$. What is consumed or broken down and not built into the deposit stays in the bath as breakdown products, counted in mL/L of the additive they came from, as shown in Equation (10.2).

$$V\frac{\mathrm{d}B}{\mathrm{d}t} = \sum_k y_k\left[c_k I\frac{C_k}{C_{k,V}} + d_k(T)\,V C_k\right] - (b+w)\,B \tag{10.2}$$

The yields are $y_S = 0.8$, $y_P = 1.0$ and $y_L = 0.5$, and no term removes B except bath taken out. At the standard load, a cycle-average current of 9.09 A, the accelerator's 1200 mL are consumed at 45.4 mL/h and broken down at 12.0 mL/h, a turnover of 20.9 h. The suppressor turns over in 8.2 days and the leveler in 3.5, and the bath forms 0.258 mL/L of breakdown products an hour, or 0.463 mL/L under a lot of thick-copper wafers.

10.3.2 Copper, Acid, Temperature and the Exchanges

The soluble anode dissolves copper at the efficiency $\eta_a = 100\,\%$ while the cathode plates it at $\eta_c = 98\,\%$, and the balance of the cathode current makes hydrogen and consumes acid, as shown in Equation (10.3).

$$\begin{aligned} V\frac{\mathrm{d}Cu}{\mathrm{d}t} &= (\eta_a - \eta_c)\,\gamma_{Cu} I + \dot{m}_{Cu} - (b+w)\,Cu + b\,Cu_V \\ V\frac{\mathrm{d}H}{\mathrm{d}t} &= -(1-\eta_c)\,\gamma_H I + \dot{m}_H - (b+w)\,H + b\,H_V \\ \tau_T\frac{\mathrm{d}T}{\mathrm{d}t} &= T_{set} + \Delta T_{ch} - T \end{aligned} \tag{10.3}$$

Here $\gamma_{Cu} = 3600\,M_{Cu}/2F = 1.1855\,\mathrm{g}$ of copper per A h, $\gamma_H = 3600\,M_{H_2SO_4}/2F = 1.8297\,\mathrm{g}$ of acid per A h, $\dot{m}_{Cu}$ and $\dot{m}_H$ are the titration additions in g/h, and $\tau_T = 30\,\mathrm{min}$. At the standard load the anode adds 5.2 g of copper a day to the 8 kg the bath holds and the hydrogen consumes 8.0 g of acid, so in no documented day does either move by more than 0.07 g/L. Two exchanges are discrete. Each wafer leaves with $v_d = 30\,\mathrm{mL}$ of bath, which the feed replaces with VMS, $X \leftarrow X + (v_d/V)(X_V - X)$ for every species X with $B_V = 0$, equivalent to $\dot{q}_d = 0.5\,\mathrm{L/h}$ at the 216 s cycle. A dump-and-remake replaces all 200 L at once and holds the tool for $t_{rm} = 120\,\mathrm{min}$. The tool's DI auto-fill replaces evaporation, so V is fixed.

10.3.3 Wafers and Faraday's Law

Each wafer is a discrete event. It enters at a multiple of the cycle $t_{cyc} = 216\,\mathrm{s}$ and runs a three-step waveform, nucleation for $t_{nuc} = 10\,\mathrm{s}$ at $I_{nuc} = 3\,\mathrm{A}$, fill for $t_{fill} = 30\,\mathrm{s}$ at $I_{fill} = 6\,\mathrm{A}$ and bulk for t_b at I_b, which is 14 A for a standard 1 µm wafer and twice that for a 2 µm thick-copper wafer. Its thickness follows from Faraday's law on the charge it received, Equation (10.4).

$$\begin{aligned} Q_w &= t_{nuc} I_{nuc} + t_{fill} I_{fill} + t_b I_b \\ h &= \frac{\eta_c\,Q_w\,M_{Cu}}{2F\,\rho_{Cu} A} = \frac{\eta_c\,Q_w}{3600\,\kappa_F} \end{aligned} \tag{10.4}$$

The charge Q_w is in A s with the times in s, the 300 mm wafer's area is $A = 706.9\,\mathrm{cm}^2$, copper's density is $\rho_{Cu} = 8.96\,\mathrm{g/cm}^3$, and $\kappa_F = 2F\rho_{Cu}A/(3600\,M_{Cu})$ is 0.534 A h per µm. A standard wafer at 98 % takes 0.545 A h, a bulk step of 125.2 s, and a thick one 1.090 A h and 132.7 s, so the cycle-average current doubles from 9.09 to 18.17 A under a thick lot. The current is held over each physics step at its exact average, so every wafer's charge is exact and the fourth-order Runge-Kutta step integrates a smooth right-hand side. A four-point probe reports each wafer's mean thickness with 0.4 % repeatability as it leaves.

10.3.4 The Fill Window

Whether a wafer's vias fill is read from the true chemistry at the middle of its fill step, through the illustrative window of Equation (10.5).

$$
\begin{aligned}
r &= \frac{S}{P\,g(Cl)}, \qquad g(Cl) = \frac{Cl}{K_{Cl}+Cl}\,\frac{K_{Cl}+Cl_V}{Cl_V} \\
r_c &= r_0\Big(\frac{L}{L_V}\Big)^{\kappa_L}, \qquad w(B) = w_0\Big(1-\frac{B}{B_{cl}}\Big) \\
m &= \min\Big[\frac{w(B)-|\ln(r/r_c)|}{w_0},\ \frac{Pg-P_{min}}{P_V-P_{min}}\Big]
\end{aligned}
\tag{10.5}
$$

The window is a band of the effective accelerator-to-suppressor ratio r, in which chloride weights the suppressor, with $g = 1$ at the make-up's 50 mg/L and $K_{Cl} = 20\,\mathrm{mg/L}$. It is centered at $r_0 = 3.0$ at the make-up leveler and moved up by more leveler with $\kappa_L = 0.5$, and its half-width in natural-log units, $w_0 = 0.35$ in a fresh bath, narrows linearly with the breakdown products until it closes at $B_{cl} = 30\,\mathrm{mL/L}$. The margin m is 1 at the center of a fresh window, 0 on its edge and negative outside it, and below $P_{min} = 0.5\,\mathrm{mL/L}$ of effective suppressor the field is not covered at any ratio. A wafer with $m < 0$ voids, as an underfill with a seam when $r < r_c$ and as a pinch-off void when $r > r_c$ or when the suppressor is short. Past B_{cl} the half-width turns negative, so the margin keeps saying how far past closing the bath is. The window, the rates of Equations (10.1) and (10.2) and the yields are illustrative teaching values rather than a supplier's data, and no feature-scale deposition is computed.

10.3.5 Starting Point, Step and Checks

Table 10.2 gives every parameter with its class. Visible parameters carry a slider in the Show Model panel, hidden ones never appear there, and fixed ones are constants of the code. Every scenario starts with every species at its make-up strength and the breakdown products at 22 mL/L, except the aged baths of cases 2 and 5 at 27 and 32 mL/L, and the controllers start at the doses and bleed that hold that bath. In case 4 they are 57.43 mL/h of accelerator, of which 55.16 mL/h is feedforward, 2.035 mL/h of suppressor, 7.197 mL/h of leveler, a bleed of 1.76 L/h and a bulk step of 125.2 s. The engine of Section 1.6.1 integrates the chemistry and the ledgers together by the fourth-order Runge-Kutta method with a constant step of 1 s, the control layer executes once per 10 s, and the historian stores a row per simulated minute. One seed, 20261102, drives all of the noise, so every analysis and probe result repeats exactly from run to run. Speeds from 1 to 1200 simulated minutes per second of wall time, with 10 as the default, change how fast the trajectory is drawn but not the trajectory. At the default speed a day takes 2.4 min of wall time, and without the interface it computes in about 0.07 s. With the step halved, the challenge's virgin make-up moves by 7×10^{-7} L from 57.0332 L and its worst fill margin by 1×10^{-7} from 0.1178, because the controller and the analyses keep their times and only the integration is refined. The species ledgers close to 3×10^{-12} relative in every case, and the application's 40 verification checks include closure below 1×10^{-9}.

10.3.6 Assumptions

The bath is one well-mixed volume of fixed size. A real tool has a reservoir, several cells and piping with mixing times of minutes, which are short against the hours in which this chemistry moves, and the concentrates' own volume is neglected.

The three additives are lumped species in mL/L of vendor concentrate, consumed per amp-hour and broken down at a first-order rate that grows exponentially with temperature. The coefficients are illustrative and chosen so that a day shows the lessons, which is why the accelerator turns over in 20.9 h. The breakdown products are one lumped species with fixed yields, although the fragments of a real accelerator or suppressor can keep part of its activity.

The fill window is an illustrative correlation. A real window is mapped by cross-sectioning filled test structures for one additive package, one waveform and one feature size, and it also depends on

Table 10.2 Parameters of the Bath and Whether Show Model Exposes Them

Symbol	Meaning	Value	Class
V, τ_T	bath volume, thermal lag	200 L, 30 min	hidden
T_{set}	chiller setpoint	25 °C	visible
S_V, P_V, L_V	additives in the VMS	6.0, 2.0, 3.0 mL/L	visible
Cl_V, Cu_V, H_V	chloride, copper, acid in VMS	50 mg/L, 40, 10 g/L	visible
t_{cyc}	wafer cycle, 400 a day	216 s	hidden
t_{nuc}, I_{nuc}	nucleation step	10 s, 3 A	visible
t_{fill}, I_{fill}	fill step	30 s, 6 A	visible
I_b	bulk current, standard wafer	14 A	visible
—	thick-copper factor, t_b limits	2, 60 s to 146 s	hidden
v_d	drag-out per wafer	30 mL	visible
η_a, η_c	anode, cathode efficiency	100, 98 %	visible
t_{rm}	dump-and-remake downtime	120 min	visible
c_k	consumption per A h, S, P, L	5.0, 0.18, 0.66 mL	visible
c_{Cl}	chloride into the deposit	0.5 mg per A h	hidden
$d_{k,25}$	breakdown at 25 °C, S, P, L	1.0, 0.1, 0.2 %/h	visible
E_k	breakdown temperature slope	0.12, 0.07, 0.07 K^{-1}	hidden
y_k	breakdown-product yields	0.8, 1.0, 0.5	hidden
r_0, w_0	window center, half-width	3.0, 0.35	visible
B_{cl}, κ_L	closing level, leveler slope	30 mL/L, 0.5	visible
K_{Cl}, P_{min}	chloride half-strength, least P	20 mg/L, 0.5 mL/L	hidden
$D_{k,max}$	pump capacities S, P, L	150, 12, 40 mL/h	hidden
b_{max}	bleed capacity	4 L/h	visible
—	CVS interval, analysis time	120, 40 min	visible
—	CVS sample line	2 min	fixed
—	CVS repeatability, S	2 % of S_V	visible
—	CVS repeatability, P, L, Cl	2, 3 %, 1 mg/L	hidden
—	chance of a bad analysis	0	hidden
—	laboratory interval	480 min	visible
—	laboratory delay	60 min	hidden
—	laboratory Cu, H, B	0.2, 0.1 g/L, 0.03 mL/L	hidden
—	probe, thermometer	0.4 %, 0.05 °C	hidden
p_V, p_D	VMS price, downtime cost	$20/L, $1500/h	visible
κ_F, γ_{Cu}, γ_H	Faraday constants	0.534, 1.1855, 1.8297	fixed

Repeatabilities are 1σ. κ_F in A h per µm, γ in g per A h.

rotation, aspect ratio and pattern density, none of which is modeled. The via cross-sections in the scene are drawn from the margin rather than computed.

Faraday's law acts on the flat wafer area with one cathode efficiency, with no edge exclusion, radial profile or terminal effect. Copper and acid follow the efficiency gap, the hydrogen side reaction and the make-up, with no chemical dissolution of the anode and no anode film. The CVS analyzer is a sampled measurement with Gaussian repeatability and without the matrix effects of a real stripping analysis, and the laboratory's analysis of the breakdown products stands in for a separate chromatographic method. All sizes, rates, prices and limits are teaching values, not a design basis.

10.4 A Window Narrowed by an Inventory

Two time scales meet in the ratio. The accelerator turns over in 20.9 h at the standard load and the suppressor in 8.2 days, so a dosing error moves the accelerator within hours while the suppressor barely moves. In case 1, where fixed pumps meet a thick-copper lot at 1 h, the accelerator is at 5.19 mL/L by 6 h and the suppressor still at 1.965 mL/L, and the ratio of 2.65 has left the window through its lower edge at 2.70. The ratio follows the accelerator.

The breakdown products set how far the ratio may wander. The half-width of the window is 41.9 % of the ratio in a fresh bath and shrinks to 9.8 % at the 22 mL/L of most cases, then to 7.3, 4.8 and 3.6 % at 24, 26 and 27 mL/L, until the window closes at 30 mL/L, as shown in Figure 10.2. The leveler moves the center rather than the width. The 300 mL overdose of case 8 raises it from 3.0 to 4.5 mL/L, which moves the center from 3.00 to 3.67 and leaves a bath held at the setpoint ratio of

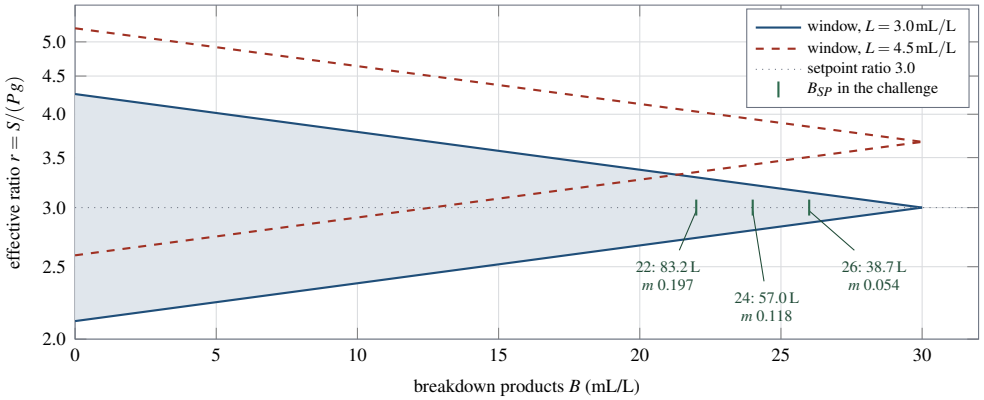


Figure 10.2 Fill window in the effective accelerator-to-suppressor ratio against the breakdown products, Equation (10.5) at the make-up chloride, on a logarithmic ratio axis. The shaded band with solid edges is the window at the make-up leveler, the dashed lines the window after the leveler overdose of case 8, and the dotted line the setpoint ratio. The vertical ticks mark the three setpoints of the breakdown products tried on the challenge day, labeled with the day's virgin make-up and the worst fill margin of its 400 wafers under the suggested dosing.

3.0 below the window at the 22 mL/L of case 8 and at every older age, as the dashed lower edge of Figure 10.2 shows.

Bleed-and-feed is the only exit the breakdown products have, and every liter of bleed is a liter of virgin make-up. Integrating Equation (10.2) over a period with no titration dilution gives the make-up balance of Equation (10.6).

$$V_{VMS} = \int_0^t (b + \dot{q}_d)\,\mathrm{d}t = \frac{\int_0^t GV\,\mathrm{d}t - V[B(t) - B(0)]}{\bar{B}_{out}} \tag{10.6}$$

Here G is the formation rate in mL/L per hour, the right-hand side of Equation (10.2) without its outflow and divided by V, and $\bar{B}_{out}$ is the flow-weighted mean concentration of the bath that left. The numerator is set by the wafers and the bath's age, the denominator by the setpoint. On the challenge day with the suggested dosing the bath forms 1868 mL of breakdown products under every bleed policy. Holding them at 22, 24 and 26 mL/L stores 28, 510 and 897 mL in the bath and removes the rest at mean exit concentrations of 22.12, 23.81 and 25.12 mL/L, which is 83.2, 57.0 and 38.7 L of virgin make-up, while the worst margins of the three days are 0.197, 0.118 and 0.054. The byproducts leave only at the concentration the bath holds them at, so each mL/L of setpoint is make-up saved and window given away.

The same balance settles the choice between bleeding and dumping. A dump cycle at the setpoint B_{SP} replaces the whole volume once every $B_{SP}/\hat{G}$ hours, which is $\hat{G}V/B_{SP}$ per hour, exactly the make-up rate at which bleed-and-feed holds B_{SP}. The application's supervisor prices a day of each policy, $24p_V(b_{hold} + \dot{q}_d)$ for the bleed against $24[(Vp_V + t_{rm}p_D)\hat{G}/B_{SP} + p_V\dot{q}_d]$ for the dump cycle, with b_{hold} the bleed that holds B_{SP}, $p_V = \$20$ per liter and $p_D = \$1500$ per hour of downtime. The dump cycle then carries the bleed's make-up plus the drag-out and the downtime, so it is never the cheaper of the two. Across 69 072 supervisor evaluations in case 2 with the bleed holding its setpoint, spanning downtime costs from 0 to \$10 000 per hour, make-up prices from 1 to \$100 per liter and setpoints from 4 to 30 mL/L, the dump cycle was never cheaper. In a bath whose byproducts accumulate linearly, a dump buys time and a younger bath rather than make-up.

🕮 **Pause.** Try Exercise 10.2 before reading on.

Time is what the supervisor can weigh. The largest bleed, $b_{max} = 4$ L/h, plus the drag-out turns the bath over with $\tau_b = V/(b_{max} + \dot{q}_d) = 44.4$ h toward $B_\infty = \hat{G}V/(b_{max} + \dot{q}_d)$. The time to bring the

Table 10.3 The Loops, Their Actions and Their Shipped Settings

Tag	CV	MV	γ	K_c	τ_I	Output
AIC-S	$\bar{y}_S$	Δ_S	+1	30	6 h	$-F_S$ to $150-F_S$
AIC-P	$\bar{y}_P$	Δ_P	+1	10	24 h	$-F_P$ to $12-F_P$
AIC-L	$\bar{y}_L$	Δ_L	+1	15	12 h	$-F_L$ to $40-F_L$
FIC-BLEED	$\hat{B}$	b	−1	0.5	12 h	0 to 4 L/h
R2R	η_{meas}	t_b	—	—	—	60 to 146 s
AIC-CU	Cu_{lab}, H_{lab}	w, $\dot{m}$	—	—	—	over 30 min

K_c in mL/h per mL/L, and in L/h per mL/L for the bleed. R2R: EWMA, $\lambda_R = 0.3$.
AIC-CU: proportional, $g_T = 0.8$. Gate $g = 0.15$, $\lambda = 0.5$.
Feedforward $\hat{c}_k$ = 4.75, 0.18, 0.66 mL per A h for S, P, L.

estimate $\hat{B}$ back to the level $B_{ok} = B_{cl}(1 - m_{min})$, at which a centered chemistry still keeps the margin m_{min}, is then given by Equation (10.7).

$$t_{rec} = \tau_b \ln \frac{\hat{B} - B_\infty}{B_{ok} - B_\infty} \tag{10.7}$$

In case 5 the bath starts at 32 mL/L, the model's formation rate of 0.248 mL/L an hour gives $B_\infty = 11.0\,\text{mL/L}$, and $m_{min} = 0.15$ gives $B_{ok} = 25.5\,\text{mL/L}$, so that $t_{rec} = 16.5\,\text{h}$ against the 2 h of a remake. The bath of case 2, at 27 mL/L, needs 4.4 h.

10.5 Implemented Controller

Six loops and a supervisor share one controller, called every $\Delta t_c = 10\,\text{s}$ or ten physics steps, although the three dosing trims change only when an analysis arrives, the titration only after a laboratory result and the run-to-run loop once per wafer. The engine step of Figure 1.11 also books the wafer events and exchanges after it integrates. Table 10.3 gives the shipped pairings and settings. Its γ is the action a of Section 1.6.2, with +1 reverse acting and −1 direct acting.

10.5.1 Instruments

No controller reads the bath. The rectifier's amp-hour counter gives the charge, whose rate over each control period passes two first-order lags of 8 min in series, so that the wafer waveform does not ripple the doses, and the result $\dot{Q}_f$ is the feedforward's load. The CVS analyzer and the laboratory are the framework's shared sampled analyzer. The CVS draws a sample at the start of the run and every 2 h after it, through a sample line that holds 2 min of bath, publishes its accelerator, suppressor, leveler and chloride 40 min later and holds them until the next result. Its repeatabilities are 2 % of the make-up strength for the accelerator and the suppressor, 3 % for the leveler and 1 mg/L for chloride. Out of service it draws nothing, loses the analysis under way and holds its last reading, which it flags as stale after two cycles, and a declared bad analysis adds an offset to its next result. The laboratory samples every 8 h and reports 1 h later, copper to 0.2 g/L, acid to 0.1 g/L and the breakdown products to 0.03 mL/L. The probe reports each wafer as it leaves, and the bath thermometer carries 0.05 °C of colored noise.

10.5.2 Amp-Hour Feedforward

Each dosing pump delivers a feedforward and a trim, as shown in Equation (10.8).

$$\begin{aligned} D_k &= \text{clamp}\big(F_k + \Delta_k,\ 0,\ D_{k,max}\big) \\ F_k &= \hat{c}_k \dot{Q}_f + d_{k,25}\, \mathrm{e}^{E_k(T_{set}-25)}\, V C_{k,SP} \end{aligned} \tag{10.8}$$

The first term of F_k is the amp-hour feedforward with the card's coefficient, and the second is the bulk breakdown the model expects at the chiller setpoint and the concentration setpoint. The suppressor

and leveler coefficients equal the bath's, while the accelerator's ships at 4.75 mL per A h, 5 % below the bath's 5.0, so that the trim has a slow model error to remove, 2.27 mL/h of the 57.43 mL/h the bath needs at the standard load. On feedforward only the trim is zero, and on feedback only there is no feedforward and the trim is the whole dose. A change of strategy moves the trim by the feedforward it adds or removes, so the dose does not jump. A pump cannot take additive out, and a trim asking for less than nothing holds its pump at zero, where it is flagged as saturated. The same split of a dose computed from the measured load and a slow feedback trim on the quality appears in the coagulant dosing of Volume 3, Chapter 6, and Section 1.4.7 places it among the feedforward paths of this volume.

10.5.3 The Sampled PI Trim

Each trim is an instance of the framework's shared PID, written in velocity form and used here without derivative action, and it executes only when an analysis arrives and passes the gate, in the form of Equation (10.9).

$$\begin{aligned} \bar{y}_n &= \lambda y_n + (1-\lambda)\,\bar{y}_{n-1}, \qquad e_n = C_{k,SP} - \bar{y}_n \\ \Sigma_n &= \Sigma_{n-1} + K_c e_n - K_c e_{n-1}, \qquad v_n = \Sigma_n + \mathscr{I}_{n-1} \\ \Delta_n &= \mathrm{clamp}\big(v_n,\ -F_k,\ D_{k,max} - F_k\big) \\ \mathscr{I}_n &= \mathscr{I}_{n-1} + \left[\frac{K_c}{\tau_I}\, e_n + \frac{\Delta_n - v_n}{\tau_I}\right] T_{s,n} \end{aligned} \tag{10.9}$$

In it, y_n denotes the nth accepted result, $\lambda = 0.5$, and $T_{s,n}$ is the time since the previous update, capped at the 120 min cycle. The law is the one of Section 1.6.2: Σ_n is the running sum of Equation (1.13), which telescopes to the positional $K_c e_n$ and is never clamped, $\mathscr{I}_n$ is the integral of Equation (1.15) with back-calculation at the tracking time τ_I, and the clamp acts on the emitted value alone. The derivative and the output rate limit are left unused, and the trim is held between analyses. Because the output is formed before the integral is advanced, the integral learned from result n first reaches the pump with result $n+1$, two hours later. In MANUAL, and on feedforward only, each trim keeps $\mathscr{I} = u_{op} - \Sigma$ at every scan, so its output always equals the value in force and nothing jumps on a return to automatic. A switch to automatic made before the run starts the day from the trims that hold the bath.

10.5.4 The Outlier Gate

Between analyses the controller integrates its own model of each additive, Equation (10.1) with the feedforward coefficients, the doses it commanded, the load $\dot{Q}_f$, the bleed, the drag-out rate and the breakdown at the measured temperature, by explicit Euler at the control period, and it keeps the history. Each result is compared with that model at the moment its sample left the bath. A channel with $|y_n - \hat{C}_k(s_n)| > g\,C_{k,SP}$ and $g = 0.15$ is rejected, and a second rejection on the same side is accepted as a real change. An accepted result re-anchors the model, $\hat{C}_k \leftarrow \hat{C}_k + y_n - \hat{C}_k(s_n)$. The gate is a gross-error test of a sparse analysis against the process model [13], and its confirmation rule costs a real step one analysis. Comparing a late reading with the model of its own time is what Section 1.4.4 asks of any reading that describes an old state of the tool.

10.5.5 Bleed and Feed

The bleed policy is off, a fixed 2.0 L/h, 0.2 L per A h, or the least bleed that holds the estimated breakdown products at B_{SP}. The last is a direct-acting PI ($\gamma = -1$) of the same law on $\hat{B}$, since more bleed is needed when the estimate is above its setpoint, as the application's Show Model note also says. Its output bias is the steady bleed $b_{ff} = \max(0, \hat{G}V/B_{SP} - \dot{q}_d)$, with the tuning of Table 10.3, and it starts at its positional output for the bath it finds. The formation rate $\hat{G}$ is Equation (10.2) evaluated on the controller's model of the additives, B_m integrates it against the bath taken out, and the laboratory

corrects the result every 8 h through the framework's shared bias update of Equation (10.10).

$$\begin{aligned} \hat{B} &= B_m + \beta \\ \beta &\leftarrow \operatorname{clamp}\big(\beta + \lambda_B[B_{lab} - B_m(s) - \beta],\ \pm 20\big) \end{aligned} \tag{10.10}$$

in which $B_m(s)$ is the model's value at the laboratory sample's time and $\lambda_B = 0.5$. Because the accelerator's coefficient is 5 % low, $\hat{G}$ runs 3.5 % low, 0.2488 against 0.2579 mL/L an hour at the reset of case 4, and the bias makes up the difference.

Pause. Try Exercise 10.3 before reading on.

10.5.6 Titration, Run-to-Run and Bath Life

After each laboratory result, copper more than 0.5 g/L above its setpoint is diluted with $g_T Ve/Cu$ liters of DI water over 30 min, copper as far below it gets $g_T Ve$ grams of copper concentrate, and acid more than 0.3 g/L low gets $g_T Ve$ grams of acid, with e the titrated error and $g_T = 0.8$. Copper rises by at most 0.042 g/L and acid falls by at most 0.065 g/L in any documented day, so these corrections never trigger in the shipped cases, and a check that starts the bath 3 g/L high in copper and 2 g/L low in acid confirms their direction. The run-to-run loop infers from each probe result the cathode efficiency that explains it, $\eta_{meas} = 3600\,\kappa_F h_{meas}/Q_w$, corrects the efficiency the recipe assumes through the shared bias update with $\lambda_R = 0.3$ [14], and sets the next bulk step from Equation (10.4), $t_b = (3600\,\kappa_F h_{SP}/\hat{\eta} - t_{nuc}I_{nuc} - t_{fill}I_{fill})/I_b$. With the true efficiency equal to the assumed one, the thickness error is the probe's 0.4 % passed through the average, $0.4\sqrt{\lambda_R/(2-\lambda_R)} = 0.168\,\%$ rms by the noise inflation of Section 1.4.3, against the 0.161 to 0.195 % the engine gives across the cases in AUTO.

The bath-life supervisor computes the margin $\hat{m}$ that the analyses and $\hat{B}$ imply through Equation (10.5), the margin a centered chemistry could still give, $m_c = \max(0, 1 - \hat{B}/B_{cl})$, and the recovery time of Equation (10.7). It advises a dump when $m_c < m_{min}$ and $t_{rec} > t_{rm}$, or when the dump cycle is cheaper and $\hat{B} \geq B_{SP}$, and it reports the reason. Set to automatic, it requests the dump once the advice has stood for 30 min, and the tool finishes the wafer in the cell, drains the bath, remakes it from VMS and stays down for the remake while the controllers' models restart at the make-up strength.

10.5.7 Simulation, Internal and Analysis Models

Section 10.3 is the simulation model. The controllers' internal models are the dosing model with its coefficients and the breakdown at the measured temperature, its formation of breakdown products, Faraday's law with the assumed efficiency and the recovery time of Equation (10.7), and only the supervisor reads the fill window. No control model is identified, and the report-to-report model of the next section is an analysis model of the trim, linearized from the simulation model.

10.6 The Gain a Two-Hour Analysis Can Carry

Around an operating point the accelerator is a first-order bath, $V\,\mathrm{d}x/\mathrm{d}t = -kx + \Delta$, where x is its deviation, Δ the trim, and $k = c_S I/S_V + d_S(T)\,V + b + \dot{q}_d$ is everything that takes accelerator out per unit of concentration. In case 3, after the new lot raises c_S to 6.25 mL per A h, $k = 14.24$ L/h at the standard load and a bleed of 2.28 L/h, the bleed a noise-free run settles to after its first day, which is a time constant $\tau = V/k$ of 14.0 h. The trim changes at report times t_n and holds for $T = 2$ h, and result n describes the bath at $s_n = t_n - \theta$ with $\theta = 42$ min. Sampling the bath at the sample times, with each trim held until the next report, gives Equation (10.11).

$$\begin{aligned} x_{n+1} &= a\,x_n + \beta_1 \Delta_{n-1} + \beta_0 \Delta_n, \qquad a = \mathrm{e}^{-T/\tau} \\ \beta_0 &= \frac{1 - \mathrm{e}^{-(T-\theta)/\tau}}{k}, \qquad \beta_1 = \mathrm{e}^{-(T-\theta)/\tau}\,\frac{1 - \mathrm{e}^{-\theta/\tau}}{k} \end{aligned} \tag{10.11}$$

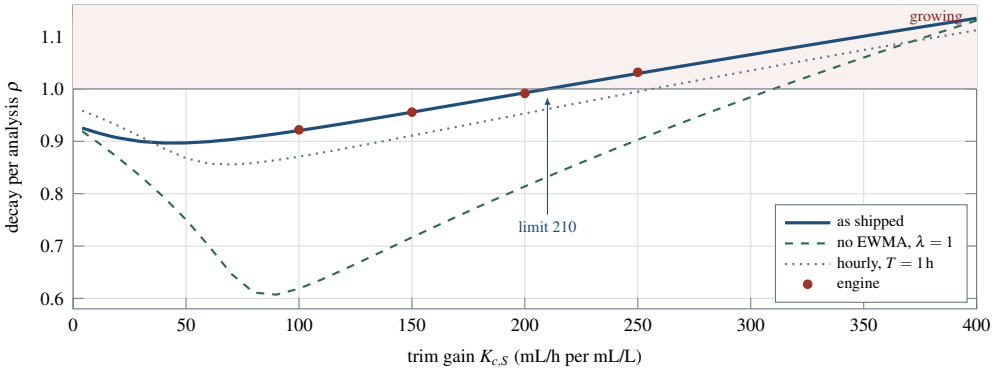


Figure 10.3 Decay per analysis of the accelerator trim against its gain, from the report-to-report recurrence at the operating point of case 3 after the new lot, with $\tau_I = 6$ h. The thick solid curve is the trim as shipped, the dashed curve the trim without the average ($\lambda = 1$) and the dotted curve the trim on an hourly analysis. Values above one grow. The round markers are noise-free runs of the engine over 96 h, with the mean half-swing ratio of the cycle-averaged accelerator converted to a factor per analysis.

The trim of Equation (10.9), unclamped and with $T_{s,n} = T$, closes the loop through $\bar{y}_n = \lambda x_n + (1-\lambda)\bar{y}_{n-1}$, $\Delta_n = -K_c\bar{y}_n + \mathscr{I}_{n-1}$ and $\mathscr{I}_n = \mathscr{I}_{n-1} - (K_cT/\tau_I)\bar{y}_n$. The result is a linear recurrence in $(x_n, \Delta_{n-1}, \bar{y}_{n-1}, \mathscr{I}_{n-1})$, and its spectral radius ρ is the factor by which an error shrinks per analysis. Section 1.4.4 derives a boundary of the same kind, Equation (1.7), for a recipe loop whose readings wait in a queue of whole runs. Here the age is a fraction of the update period, and the average and the integral add states.

At the shipped $K_c = 30$ and $\tau_I = 6$ h the recurrence gives $\rho = 0.900$, an error that shrinks by a factor of e in 18.9 h while it swings with a period of 37.6 h, and the fastest decay at this integral time is $\rho = 0.897$ at $K_c = 44$. The radius reaches one at $K_c = 210$, as shown in Figure 10.3. Noise-free runs of case 3 over 96 h confirm the boundary. Averaged over each wafer cycle to remove the plating ripple, the engine's successive half-swings of the accelerator, after the first, shrink or grow by geometric-mean factors of 0.691, 0.847, 0.973 and 1.094 at gains of 100, 150, 200 and 250, which is ρ of 0.922, 0.956, 0.992 and 1.032 against the recurrence's 0.921, 0.956, 0.993 and 1.029, with periods of 18.23, 14.67, 12.86 and 11.46 h against 18.28, 14.72, 12.74 and 11.44 h. At 300 and 400 the growing swing drives the pump to 0 and 150 mL/h and settles into a limit cycle.

The limit belongs to the measurement rather than to the chemistry. With the analysis age cut to the 2 min of the sample line the recurrence allows $K_c = 744$, at 21 min 440, and an hour of age lowers it to 141. The average costs almost a third of the gain, because without it ($\lambda = 1$) the limit is 310, while $\lambda = 0.3$ lowers it to 107. An hourly analysis with the same age raises the limit only to 256, which noise-free runs bracket between 230, where half-swings shrink by 0.913, and 280, where they grow by 1.079. The integral time moves it most, since at $\tau_I = 3$ h the loop is unstable above $K_c = 33$ and at 12 h above 307. Advancing the integral in the same report as its error, rather than one report later, would lower the limit from 210 to 198, because the integral would then add K_cT/τ_I to the immediate correction. However well the analyzer measures, a trim updated every two hours on a result 42 minutes old needs most of a day to remove an error, while a thick-copper lot doubles the consumption within the few minutes of one wafer. A continuous composition loop loses its margin in the same way behind a lagging online analyzer in Volume 5, Chapter 3.

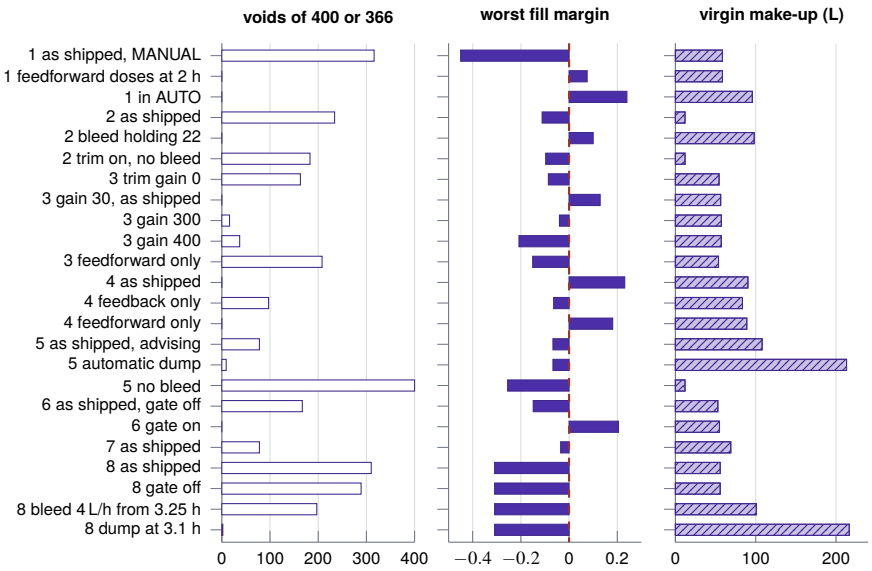


Figure 10.4 The scorecard of Table 10.5 drawn as bars: the voids of each day (open bars), the worst fill margin (solid bars) with the void edge at zero dashed, and the virgin make-up used (hatched bars). Notice that every row with a negative worst margin has voids, that more trim gain in case 3 and the gate in case 8 cost wafers, and that the two dumps buy their few voids with the largest make-up of the table.

Table 10.4 The Case Sequence

#	Scenario and event	Teaching point
1	MANUAL, thick-copper lot at 1 h	dose by the charge
2	Feedforward only, no bleed, $B = 27$	a window closes
3	Accelerator lot, c_S 5.0 to 6.25 at 1 h	the trim's gain limit
4	Thick-copper lot at 4 h	feedforward or feedback
5	$B = 32$, bleed toward 22	past the end of bath life
6	Result 50 % high at 2:40, gate off	a bad analysis
7	CVS down 6–12 h, chiller +10 °C	feedforward by force
8	+300 mL of leveler at 3 h	a dose that stays
9	400 wafers, lot at 8 h, CVS down 14–20 h	graded day

10.7 The Cases as Measured

Eight cases and a graded day make up the sequence of Table 10.4, with the briefings, completion rules and challenge of Section 1.6.4. Every case begins from the reference configuration of Table 10.3, with the feedforward and the trim, the gate, the bleed holding 22 mL/L, the titration and the run-to-run loop on and the supervisor advising, and departs from it only as its row says. Every number that follows comes from the engine at the default seed, over each scenario's 24 h and without the interface, with each setting changed before the run starts unless a time is given. Table 10.5 gathers the scorecard of each case and variant, and Figure 10.4 draws it as bars. A count of voids rests on one realization of the analysis noise wherever the ratio grazes a window edge for hours. Across ten other seeds case 4 on feedback only voids 51 to 117 wafers, case 8 voids 239 to 349 and case 7 none to 125, while the suggested configuration of the graded day plates every wafer at every seed, with worst margins of 0.083 to 0.176 and 56.5 to 57.3 L of make-up.

Table 10.5 Scorecard From the Truth Log, per Case and Variant, Over 24 h

Case and configuration	Voids	First (h)	Worst margin	VMS (L)	Wafers
1 as shipped, MANUAL	316	5.09	−0.450	58.6	400
1 feedforward doses at 2 h	0	—	0.075	58.6	400
1 in AUTO	0	—	0.240	95.8	400
2 as shipped	234	10.01	−0.112	12.0	400
2 bleed holding 22	0	—	0.100	98.1	400
2 trim on, no bleed	183	13.07	−0.097	12.0	400
3 trim gain 0	163	14.27	−0.086	54.4	400
3 gain 30, as shipped	0	—	0.129	56.4	400
3 gain 300	16	22.43	−0.040	57.2	400
3 gain 400	37	21.41	−0.208	57.1	400
3 feedforward only	208	11.57	−0.151	53.7	400
4 as shipped	0	—	0.231	90.3	400
4 feedback only	97	8.09	−0.064	83.5	400
4 feedforward only	0	—	0.181	88.9	400
5 as shipped, advising	78	0.05	−0.067	108.0	400
5 automatic dump	9	0.05	−0.067	213.1	366
5 no bleed	400	0.05	−0.255	12.0	400
6 as shipped, gate off	167	5.69	−0.149	52.9	400
6 gate on	0	—	0.205	54.9	400
7 as shipped	78	12.65	−0.035	69.0	400
8 as shipped	310	3.05	−0.309	55.8	400
8 gate off	289	3.05	−0.309	55.8	400
8 bleed 4 L/h from 3.25 h	197	3.05	−0.309	100.7	400
8 dump at 3.1 h	2	3.05	−0.309	216.5	366

10.7.1 Fixed Pumps Through a Lot Change

Case 1 runs in MANUAL with the pumps at the doses that hold the bath at the standard load, 57.4 mL/h of accelerator, 2.04 of suppressor and 7.20 of leveler, and a bleed of 1.94 L/h. At 1 h a thick-copper lot starts and 383 of the day's 400 wafers take twice the charge. The first void, a seam, comes at 5.1 h in wafer 85, the ratio reaches 2.20 by 24 h, and 316 wafers void. The manual bulk step, set for a standard wafer, also leaves every thick wafer 5.35 % thin, because the nucleation and fill steps do not double with the bulk current. Setting the pumps to the feedforward's doses at 2 h, 93.2 mL/h of accelerator, voids no wafer, although with nothing to remove the feedforward's 5 % error the accelerator drifts to 5.75 mL/L by 24 h and the worst margin is 0.075. In AUTO no wafer voids, the worst margin is 0.240, and the run-to-run loop holds every wafer within 0.65 % of its target. The case's own route to AUTO, a Reset that keeps the manual pumps at the feedforward's doses and a switch before the run, gives the same day.

10.7.2 Feedforward Alone in an Aged Bath

In case 2 the feedforward doses alone with the bleed off, from 27 mL/L of breakdown products. The additives hold, the accelerator at 5.84 mL/L and the suppressor at 2.000 mL/L after 24 h, but the drag-out's 12 L a day is the only exit and the breakdown products climb to 31.4 mL/L. The window's half-width falls from 3.6 % to 1.3 % by 10 h and is gone at 16.3 h. The first void comes at 10.0 h in wafer 167, where the feedforward's drift, a ratio 1.4 % low from the accelerator's coefficient, crosses the narrowed edge, and 234 wafers void, while the supervisor advises a dump from the start. With the bleed holding 22 mL/L it starts at its 4 L/h limit, the breakdown products fall to 21.4 mL/L by 24 h, no wafer voids, and the day uses 98.1 L of virgin make-up instead of 12. With the trim on and no bleed, 183 wafers void from 13.1 h.

10.7.3 How Much Trim a Sampled Loop Takes

At 1 h in case 3 a new lot of accelerator concentrate raises its consumption by 25 %, which the feedforward's coefficient does not know. With no trim the accelerator drifts to 5.34 mL/L and 163 wafers void from 14.3 h, and on feedforward only 208 void from 11.6 h. At the shipped gain of 30 the accelerator stays between 5.76 and 6.00 mL/L, lowest at 10.7 h, the trim reaches 10.5 mL/h by 12 h and 18.7 mL/h by 24 h against the 13.6 mL/h the new lot needs at the setpoint, and the worst margin is 0.129. A gain of 10 is slower, reaching its lowest 5.61 mL/L at 14.7 h. A gain of 100 holds the accelerator between 5.86 and 6.03 mL/L, while 200 swings it from 5.66 to 6.18 mL/L, as the recurrence of Section 10.6 expects of a loop near its limit. At 300 the swing grows into 16 voids from 22.4 h, and at 400 the accelerator swings between 6.53 mL/L at 16.7 h and 5.12 mL/L at 22.7 h and 37 wafers void from 21.4 h.

10.7.4 Feedforward Against Feedback

In case 4 a thick-copper lot starts at 4 h and 333 of the day's wafers are thick. With the feedforward and the trim the accelerator dose rises from 57.4 mL/h at 4 h to 101.8 mL/h at 6 h as the filtered amp-hours double, the accelerator stays between 5.91 and 6.01 mL/L, and no wafer voids, with a worst margin of 0.231. The bleed rises from 1.79 to 3.56 L/h over the same two hours to hold the breakdown products, and the day uses 90.3 L of make-up. On feedback only the result reported at 4:40 reads 5.96 mL/L, and the first to show the lot is reported at 6:40, of a sample drawn at 6:00. The trim then moves the dose from 58.1 to 62.6, 71.6, 81.9 and 90.2 mL/h at successive reports, about 10 mL/h a report once the average has caught up, while the accelerator falls to 5.06 mL/L at 12.6 h, and 97 wafers void from 8.1 h, in wafer 135, to 13.8 h. The run at 11.00 h is shown in Figure 10.5. On feedforward only the accelerator drifts to 5.80 mL/L by 24 h without a void, which is the slow error the trim exists to remove.

> **TRY IT IN THE APP** **4 · A Lot Mix Change**
>
> Load the scenario and press Run. Watch the accelerator pump's dose on the AIC-S card once the thick-copper lot starts at 4 h. The amp-hour feedforward lifts it from 57.4 mL/h to 101.8 mL/h by 6 h, and no wafer of the day voids.

10.7.5 Past the End of Bath Life

Case 5 starts at 32 mL/L of breakdown products, past the closing of the window, with the bleed holding toward 22 mL/L and the supervisor advising. The bleed starts at its 4 L/h maximum, the breakdown products fall to 29.4 mL/L by 6 h and 23.4 by 24 h, the window reopens when they pass 30 mL/L at 4.6 h, and 78 wafers void, pinched off, the last at 4.7 h. The bleed stays within 0.05 L/h of its maximum all day, because the estimate never comes down to its setpoint, so the 108.0 L of make-up are 96.0 L of bleed and 12.0 of drag-out. The supervisor reports from the first minutes that the largest bleed needs 16.5 h to restore the margin. On automatic it requests the dump at 0.50 h and the tool drains the bath at 0.53 h, once its wafer is out, so 9 wafers void before the dump and none after, with 366 wafers plated instead of 400 and 213.1 L of make-up, 200 L of it the remake. The remade bath climbs only to 5.51 mL/L by 24 h, with the bleed shut. With no bleed at all every wafer voids and the breakdown products reach 36.1 mL/L. No setting of the bleed reopens the window in time, and the application says so.

10.7.6 A Bad Analysis and the Gate

In case 6 the result reported at 2:40 reads 9.04 mL/L for a bath at 6.00, with the gate off. The trim takes it at face value and cuts the accelerator dose from 58.2 to 12.3 mL/h. At 4:40 the average still carries half the bad value, 6.56 mL/L against a bath at 5.58, so the dose recovers only to 26.0 mL/h, the accelerator falls to 5.19 mL/L at 10.7 h, and 167 wafers void from 5.7 to 15.6 h. With the gate on,

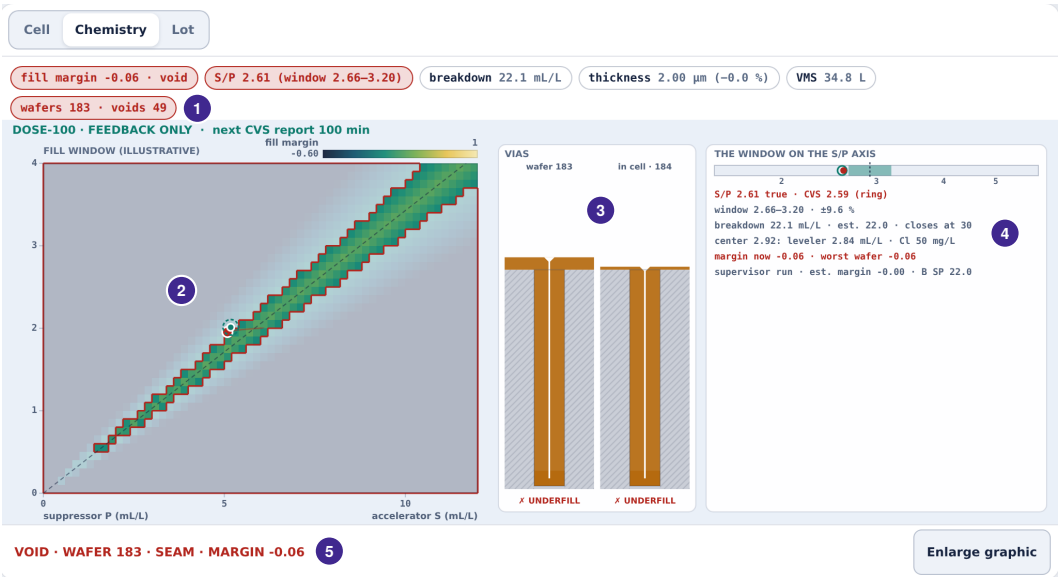


Figure 10.5 The bath in case 4, a lot mix change, at $t = 11.00$ h with the dosing strategy on feedback only, seven hours after the thick-copper lot started, in the Chemistry view, cropped to the plant view. (1) The badges read wafers 183, voids 49 and 34.8 L of virgin make-up used. (2) The fill margin baked over accelerator and suppressor at the bath's leveler, chloride and breakdown products shows the window as a narrow wedge from the origin with a red outline, the true operating point as a filled dot, red because its margin is negative, with its trail just below the wedge, and the last CVS result as a ring. (3) Both via cross-sections, wafer 183 and wafer 184 in the cell, are drawn underfilled with a seam. (4) The window on the S/P axis gives the true ratio 2.61 against a window of 2.66 to 3.20 (±9.6 %) centered at 2.92 by the leveler at 2.84 mL/L, the CVS ratio of 2.59, corrected for the analyzed chloride, as the ring, breakdown products of 22.1 mL/L (estimate 22.0), a margin now of −0.06 and a worst wafer of −0.06. (5) The status line reads VOID · WAFER 183 · SEAM · MARGIN −0.06.

the model predicts 5.96 mL/L for that sample, the result is rejected, and no wafer voids. The gate rejects nothing else in the day's 12 reports.

> **TRY IT IN THE APP** **6 · An Outlier and the Gate**
>
> Press Run with the gate off, as shipped, and watch the accelerator pump after the result of 2:40, which reads 9.04 mL/L for a bath at 6.00. The trim cuts the dose from 58.2 to 12.3 mL/h, and 167 wafers void. Press Reset, set Outlier gate to On on the DOSE-100 card before Run, and run again: no wafer voids.

10.7.7 The Analyzer Out of Service

In case 7 the CVS is out of service from 6 to 12 h, its last result reported at 4:40, and at 6.5 h the chiller drifts so that the bath reaches 34.5 °C by 8 h. At 35 °C the accelerator breaks down 3.32 times as fast as at 25 °C, an extra 2.3 % of it an hour, and the time-based feedforward still assumes the setpoint. The dose stays between 58.9 and 59.5 mL/h and the reading at 5.89 mL/L while the accelerator falls from 6.02 at 6 h to 5.50 mL/L at 12 h. The controller's own model, which uses the measured temperature, follows the fall from 5.93, where the 4:40 result left it, to 5.50 mL/L at 12 h, but no dose is computed from it. The first void comes at 12.6 h in wafer 211, a minute before the returning analyzer's first result at 12:40, of a sample drawn at 12:00. The trim then climbs from 4.0

Table 10.6 Configurations Tried on the Graded Day

Configuration	Voids	First (h)	Worst margin	VMS (L)	B at 24 h
as shipped	226	10.49	−0.374	12.0	28.86
suggested[a]	0	—	0.118	57.0	24.55
suggested, $B_{SP} = 22$	0	—	0.197	83.2	22.14
suggested, $B_{SP} = 26$	0	—	0.054	38.7	26.48
suggested, bleed 4 L/h	0	—	0.260	108.0	20.25
feedforward and trim, no bleed	60	20.45	−0.059	12.0	29.81
feedback only, $B_{SP} = 24$	216	11.09	−0.235	49.5	24.53

[a] feedforward and gated trim, bleed holding $B_{SP} = 24\,\text{mL/L}$

to 39.8 mL/h as it learns the hot bath, the accelerator bottoms at 5.37 mL/L at 14.7 h, and 78 wafers void, the last at 17.3 h.

10.7.8 A Dose That Cannot Be Taken Back

In case 8 the leveler pump puts 300 mL into the bath at 3 h, from 3.0 to 4.5 mL/L, and the window's center moves from 3.00 to 3.67 with it, so the bath's ratio of 3.0 lies below the window and the first void comes at 3.05 h in wafer 51. The gate rejects the first result to show the overdose, reported at 4:40, and accepts the second, at 6:40, on the same side. From then on the leveler pump sits at or near zero against its lower limit, the leveler leaves only by consumption, breakdown and the bath taken out, falling to 3.41 mL/L by 24 h, and 310 wafers void until 21.6 h. With the gate off 289 void until 20.3 h, so the gate's confirmation costs 21 wafers here. A fixed bleed of 4 L/h from 3.25 h ends the voids at 14.8 h with 197 voids and 100.7 L of make-up, and a dump at 3.1 h ends them at once, with 2 voids, 366 wafers and 216.5 L. Raising the accelerator setpoint to 7.2 mL/L from 3.25 h to chase the moved center still voids 251 wafers, 175 of them pinched off, because the window moves back as the leveler falls.

> ✦ DID YOU KNOW? **One Gate, Two Verdicts**
>
> From the simulator: in Table 10.5 the outlier gate takes case 6 from 167 voids to none. In case 8 the same gate lets 310 wafers void against 289 without it, because it waits for a second result before it believes the leveler overdose.

10.7.9 The Graded Day

The challenge runs 24 h from a bath at 22 mL/L, with a thick-copper lot from 8 h to the end and the CVS out of service from 14 to 20 h, and it asks for 400 wafers with no void and at most 65 L of virgin make-up. The day counts only on the bath as given, without a changed model parameter or disturbance. It ships on feedback only with no bleed, and Table 10.6 sets the measured alternatives side by side.

As shipped, 226 wafers void, the first at 10.5 h in wafer 175, the accelerator falls to 4.63 mL/L at 20.7 h, the breakdown products reach 28.9 mL/L, and only the 12 L that replace the drag-out are used. The suggested configuration, the feedforward with the gated trim and the bleed holding 24 mL/L, plates all 400 without a void, with the accelerator between 5.82 and 6.03 mL/L and a worst margin of 0.118, while the breakdown products rise to 24.6 mL/L as the bleed climbs to 2.9 L/h under the thick lot, and it uses 57.0 L, 8.0 L inside the ceiling. Each element pays for itself. The feedforward and trim without a bleed void 60 wafers from 20.4 h, when the breakdown products have reached 28.5 mL/L, and feedback only with the bleed at 24 voids 216 from 11.1 h. The setpoint prices the window as Equation (10.6) accounts for it, 83.2 L at 22, 38.7 at 26 with a worst margin of 0.054, and 108.0 with the bleed at its maximum.

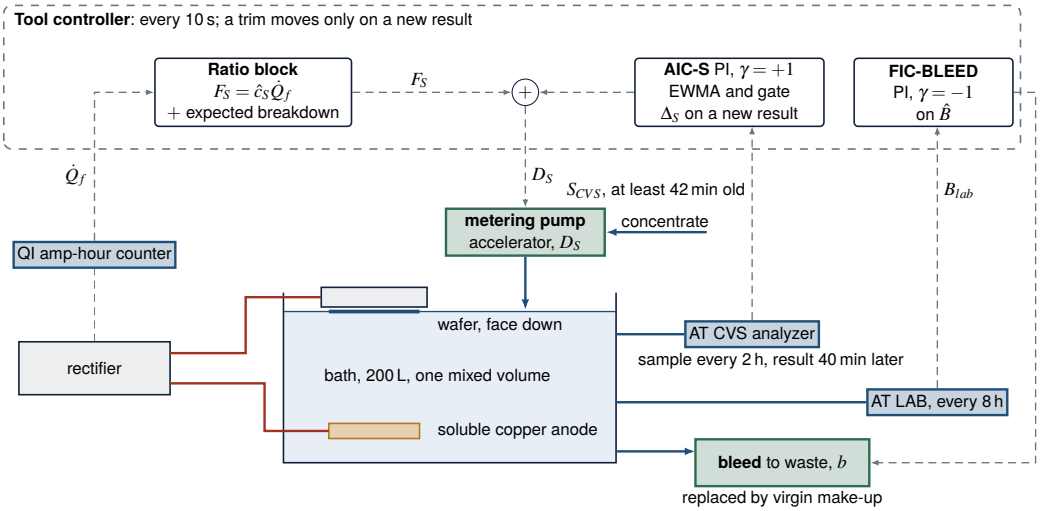


Figure 10.6 The accelerator loop and the bleed of the simulated bath drawn with their hardware. The amp-hour counter QI on the rectifier gives the load, and the ratio block turns it into the feedforward dose. The CVS analyzer AT draws through its sample line every 2 h and reports 40 min later, and AIC-S, a velocity-form PI with $\gamma = +1$, adds its trim only when a result passes the gate. The sum drives the metering pump. The laboratory corrects the model's estimate of the breakdown products every 8 h, and FIC-BLEED, direct acting with $\gamma = -1$, sets the bleed that virgin make-up replaces. Notice that the only signal that reaches the pump within minutes is the charge, not the chemistry.

10.8 Technology in Practice

The loops of Section 10.5 run on a stripping analyzer and a laboratory, the rectifier's amp-hour counter, three metering pumps and a bleed, and the controller of the plating tool. Figure 10.6 draws the accelerator loop and the bleed with that hardware.

What Measures It

The model of Section 10.5 reduces the analyzer to a sample drawn every 2 h through a 2 min line, a result published 40 min later and held, and a repeatability of 2 % of the make-up strength for the accelerator, with no matrix effects.

> **TECHNOLOGY Sensor: A Cyclic Voltammetric Stripping Analyzer**
>
> The analyzer sweeps the potential of a rotating inert disk back and forth in a sample of the bath, so that copper is plated on the disk and stripped off again on every cycle. The charge that strips the copper follows the rate of plating, which the additives set, and the method rests on that effect [16]. A result is therefore the strength of the sample's action on the deposit, expressed as a concentration. A direct analysis of the organic compounds of the bath is not commonly available by such electrochemical techniques [17]. A bad analysis shows on the trend as a single step in the reading that the next result does not repeat, as in case 6.

No open, vendor-neutral source gives a production instrument's analysis time or accuracy, so the model's values are teaching values. The one signal of consumption that arrives at once is the rectifier's charge, passed through two lags of 8 min. The laboratory's analysis of the breakdown products stands for a separate chromatographic method.

What Moves It

Each pump is modeled as delivering exactly the dose it is sent, held between zero and 150, 12 and 40 mL/h for the accelerator, suppressor and leveler, and the bleed between 0 and 4 L/h.

TECHNOLOGY Final Element: A Metering Pump on a Current Signal

A metering pump packs its pump mechanism, motor and drive electronics in one unit. A 4 to 20 mA signal sets its flow from zero to full scale, and a knob on the pump sets the flow that full scale means. Many such pumps drive a plunger with a solenoid at a variable rate, so the flow arrives as discrete pulses, a simple, reliable design where an uneven flow is acceptable. In water treatment each chemical is proportioned to the flow of the water it treats, so its pump must be set by the control system, not by hand [18]. A pump that runs away, as in case 8, puts in what no pump can take out.

The model's exact dose stands for the average such a pump delivers, with no pulses, no calibration drift and no failure but the runaway of case 8. Its clamp at zero is the one-sided actuator of that case.

Where the Controller Runs

In the app the six loops and the supervisor share one controller. In a fab a run-to-run correction such as that of the bulk step belongs to advanced process control, which combines model-based control with fault detection and classification [19]. Supervisory and lower-level loops run with cycle times from milliseconds to minutes [20]. The app integrates every 1 s and executes its controller every 10 s, while each trim moves only when an analysis arrives, every 2 h, the bulk step once per wafer, every 216 s, and the estimate of the breakdown products every 8 h.

IN PRACTICE How the Bath Is Analyzed

A review of damascene plating management names bath concentration measurement and control, film metrology, contamination and defects as the areas such a process must measure [21]. In a stripping analysis the suppressor is found by adding the sample step by step to virgin make-up solution until the stripping charge falls to half, the accelerator by adding known standards, and the leveler on a calibration curve. Copper, acid and chloride are titrated or read by absorbance. The analyst holds every solution at one temperature and checks the platinum electrode, which the additives foul over many analyses [22].

In such a controller the chapter's structure is a ratio block on the filtered amp-hour rate, three PI blocks executed on a new analysis, a bath model that gates each result, a bias update and a PI for the bleed, an EWMA on the bulk step, and the supervisor.

What Is Modernizing It

The first development is analysis at the tool. A microelectrode gave the stripping analysis without a rotating disk, for monitoring in the bath itself, in 2007 [23], and in 2025 a portable microfluidic workstation detected the additives online with a mean relative error below 10 %, replacing the laboratory's manual analysis [24]. Chemometrics serve the electroanalytical control of bath concentrations [21]. The second is dosing from a model. Since an exact analysis needs equipment a plating plant typically cannot fully provide, one approach models the electrolyte's composition from data the plant's production systems already record and doses from it, validated on an industrial zinc-nickel line [25]. This chapter's controller already integrates such a model between analyses, but no loop yet doses from it (Section 10.9).

10.9 Limitations

The fill window is an illustrative correlation in the ratio, the leveler and the breakdown products, not a supplier's data or a feature-scale deposition model, and the via cross-sections are drawn from the margin. The breakdown products are one lumped species with fixed yields, the additives' rates are chosen so that a day shows the lessons, the CVS has Gaussian repeatability and no matrix effects, the bath is one well-mixed volume, and Faraday's law acts on a flat wafer with one efficiency.

OPEN CHALLENGE Seeing What the Additives Become

Total organic carbon was once used to follow the decomposition of a plating bath's organics, but it did not correlate strongly with the bath's effective composition. Extraction followed by liquid chromatography has since identified and quantified the byproducts in a sample of a live bath [17]. The breakdown products of this chapter set the width of the fill window, yet its laboratory reports them only every eight hours. Progress would be a measurement of them fast enough for the bleed to act on, in place of an estimate corrected three times a day.

Two parts of the control system are present but idle in these cases. The copper and acid corrections never trigger, because neither species leaves its dead band within a day, and the supervisor's cost comparison cannot choose a dump in a bath whose byproducts accumulate linearly, so every dump the application calls rests on feasibility. The controller's own model tracks the accelerator through an outage, but no loop doses from it, so the drift of case 7 goes uncorrected although the information to correct it exists. All sizes, rates, prices and limits are teaching values.

10.10 Conclusion

An acid copper plating bath is a chemistry dosed against a rare analysis, and the simulator of this chapter logs each wafer's true fill margin beside the results its controllers saw. At the standard load the accelerator turns over in 20.9 h and the suppressor in 8.2 days, so the ratio follows the accelerator, and the breakdown products narrow the window to 9.8 % at 22 mL/L and close it at 30. The amp-hour feedforward answers a doubled load at once, while feedback alone first sees it at 6:40, two hours and forty minutes after the lot, and voids 97 wafers.

A report-to-report model places the gain limit of the sampled trim at 210 mL/h per mL/L, set by the analysis's 42 min age and its average, and the engine's swings decay below it and grow above it as predicted. The make-up a bleed spends is the breakdown products removed divided by the concentration at which they leave, 57.0 L at 24 mL/L on the graded day, and a bath past its life is recovered only by a dump, which ends the voids of case 5 after 9 wafers at a cost of 34 wafers.

The window and the rates are illustrative. A fuller model would add a feature-scale fill model, matrix effects in the analysis, fragments that stay partly active, and a dose computed from the controller's own model between analyses. The bath shows that a sampled analysis can correct a model but cannot carry a load.

KEY POINTS

- A copper bath uses up its additives with every amp-hour, and its analysis arrives every two hours to describe the bath of at least 42 min before.
- The amp-hour feedforward answers a doubled load at once, while feedback alone first sees it at 6:40 and voids 97 wafers.
- The trim's gain limit, 210 mL/h per mL/L, is set by the analysis's age and its average, not by the chemistry.
- The make-up a bleed spends is the breakdown products removed divided by their exit concentration, and only a dump recovers a bath past its life.

Where else this applies. A dose computed from the load with a slow quality trim in Volume 3, Chapter 6, a loop behind a lagging analyzer in Volume 5, Chapter 3, and a recipe corrected on late results in Chapter 6.

References

[1] D. Edelstein, J. Heidenreich, R. Goldblatt, W. Cote, C. Uzoh, N. Lustig, P. Roper, T. McDevitt, W. Motsiff, A. Simon, J. Dukovic, R. Wachnik, H. Rathore, R. Schulz, L. Su, S. Luce, and J. Slattery, "Full copper wiring in a sub-0.25 µm CMOS ULSI technology," in *International Electron Devices Meeting. IEDM Technical Digest*, Washington, DC, USA, 1997, pp. 773–776.

[2] P. C. Andricacos, C. Uzoh, J. O. Dukovic, J. Horkans, and H. Deligianni, "Damascene copper electroplating for chip interconnections," *IBM Journal of Research and Development*, vol. 42, no. 5, pp. 567–574, 1998.

[3] J. P. Healy, D. Pletcher, and M. Goodenough, "The chemistry of the additives in an acid copper electroplating bath," *Journal of Electroanalytical Chemistry*, vol. 338, no. 1-2, pp. 155–165, 1992.

[4] T. P. Moffat, D. Wheeler, M. D. Edelstein, and D. Josell, "Superconformal film growth: Mechanism and quantification," *IBM Journal of Research and Development*, vol. 49, no. 1, pp. 19–36, 2005.

[5] P. M. Vereecken, R. A. Binstead, H. Deligianni, and P. C. Andricacos, "The chemistry of additives in damascene copper plating," *IBM Journal of Research and Development*, vol. 49, no. 1, pp. 3–18, 2005.

[6] A. C. West, "Theory of filling of high-aspect ratio trenches and vias in presence of additives," *Journal of The Electrochemical Society*, vol. 147, no. 1, p. 227, 2000.

[7] A. Frank and A. J. Bard, "The decomposition of the sulfonate additive sulfopropyl sulfonate in acid copper electroplating chemistries," *Journal of The Electrochemical Society*, vol. 150, no. 4, p. C244, 2003.

[8] C.-K. Hu and J. M. E. Harper, "Copper interconnections and reliability," *Materials Chemistry and Physics*, vol. 52, no. 1, pp. 5–16, 1998.

[9] D. Tench and J. White, "Cyclic pulse voltammetric stripping analysis of acid copper plating baths," *Journal of The Electrochemical Society*, vol. 132, no. 4, pp. 831–834, 1985.

[10] R. D. Gudi, S. L. Shah, and M. R. Gray, "Multirate state and parameter estimation in an antibiotic fermentation with delayed measurements," *Biotechnology and Bioengineering*, vol. 44, no. 11, pp. 1271–1278, 1994.

[11] J. H. Lee and M. Morari, "Robust inferential control of multi-rate sampled-data systems," *Chemical Engineering Science*, vol. 47, no. 4, pp. 865–885, 1992.

[12] J. J. Downs and E. F. Vogel, "A plant-wide industrial process control problem," *Computers & Chemical Engineering*, vol. 17, no. 3, pp. 245–255, 1993.

[13] R. S. H. Mah and A. C. Tamhane, "Detection of gross errors in process data," *AIChE Journal*, vol. 28, no. 5, pp. 828–830, 1982.

[14] A. Ingolfsson and E. Sachs, "Stability and sensitivity of an EWMA controller," *Journal of Quality Technology*, vol. 25, no. 4, pp. 271–287, 1993.

[15] D. E. Seborg, T. F. Edgar, D. A. Mellichamp, and F. J. Doyle, III, *Process Dynamics and Control*, 4th ed. Hoboken, NJ: John Wiley & Sons, 2016.

[16] D. Tench and C. Ogden, "A new voltammetric stripping method applied to the determination of the brightener concentration in copper pyrophosphate plating baths," *Journal of The Electrochemical Society*, vol. 125, no. 2, pp. 194–198, 1978.

[17] L. D'Urzo, H. Wang, A. Pa, and C. Zhi, "SPE-HPLC detection of organic additives in acidic copper plating baths," *Journal of The Electrochemical Society*, vol. 152, no. 4, p. C243, 2005.

[18] T. R. Kuphaldt, *Lessons in Industrial Instrumentation*, version 2.33 ed. Open textbook, Creative Commons Attribution 4.0 license, ibiblio.org/kuphaldt/socratic/sinst, 2024.

[19] A. V. Prabhu, "Performance monitoring of run-to-run control systems used in semiconductor manufacturing," Ph.D. dissertation, The University of Texas at Austin, 2008. [Online]. Available: http://hdl.handle.net/2152/17721

[20] K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, A. Hahn, S. Saravia, A. Sherule, and M. Thompson, "Guide to operational technology (OT) security," National Institute of Standards and Technology, Gaithersburg, MD, Special Publication 800-82r3, 2023.

[21] R. Carpio and A. Jaworski, "Review—management of copper damascene plating," *Journal of The Electrochemical Society*, vol. 166, no. 1, pp. D3072–D3096, 2019.

[22] K. Ichiba and K. Watanabe, "Cyclic voltammetric stripping (CVS) for the measurement of additives in electro copper plating baths," *Denki Kagaku*, vol. 88, no. 2, pp. 162–165, 2020, in Japanese.

[23] B.-G. Xie, J.-J. Sun, X.-B. Chen, J.-H. Chen, T.-L. Xiang, and G.-N. Chen, "In situ monitoring of additives in copper plating baths by cyclic voltammetric stripping with a microelectrode," *Journal of The Electrochemical Society*, vol. 154, no. 10, p. D516, 2007.

[24] Y. Zhao, J.-X. Zeng, J.-Q. Yang, T. Song, R. Hu, J.-J. Su, B. Zhang, F.-Z. Yang, D. Zhan, and L. Han, "On-line detection of additive concentrations in acidic copper plating solution for metal interconnection by an electrochemical microfluidic workstation," *Industrial Chemistry & Materials*, vol. 3, no. 5, pp. 607–617, 2025.

[25] A. Leiden, S. Kölle, S. Thiede, K. Schmid, M. Metzner, and C. Herrmann, "Model-based analysis, control and dosing of electroplating electrolytes," *The International Journal of Advanced Manufacturing Technology*, vol. 111, no. 5-6, pp. 1751–1766, 2020.

10.11 Exercises

10.1 Concept check · 2 min

In scenario 4 on feedback only, the thick-copper lot starts at 4:00, yet the accelerator result reported at 4:40 reads 5.96 mL/L, close to its setpoint of 6.00. Why does that result not show the lot?

(a) The outlier gate rejected the low result and held the last good one
(b) Its sample left the bath just before the first thick-copper wafer plated
(c) The EWMA weight of 0.5 hides half of every change in a new result
(d) The accelerator does not fall until the breakdown products start to rise

10.2 Concept check · 1 min

Section 10.4 finds that in this bath a dump cycle at a setpoint B_{SP} never costs less than bleed-and-feed holding the same B_{SP}. Which reason is the one the make-up balance gives?

(a) A dump throws away the whole accelerator inventory, which costs more than the make-up
(b) The supervisor charges every void plated during the remake to the dump
(c) A dump cycle spends the bleed's make-up per hour and adds the downtime
(d) The bleed removes the breakdown products at a higher concentration than a dump does

10.3 Concept check · 2 min

In the convention of Section 1.6.2, how is the PI that sets the bleed from the estimated breakdown products $\hat{B}$ classified?

(a) Reverse acting, action $+1$, because the bleed lowers the breakdown products it controls
(b) Reverse acting, action $+1$, because its output bias is the steady bleed b_{ff}
(c) Neither, because the bleed is a feedforward of $\hat{G}V/B_{SP}$ with no feedback on $\hat{B}$
(d) Direct acting, action -1, because the bleed must rise when $\hat{B}$ is above its setpoint

10.4 Theory and calculation · 7 min

Use Equation (10.4) with $F = 96485\,\mathrm{C/mol}$, $M_{Cu} = 63.546\,\mathrm{g/mol}$, $\rho_{Cu} = 8.96\,\mathrm{g/cm^3}$, $A = \pi(15\,\mathrm{cm})^2$, $\eta_c = 0.98$ and the waveform of Table 10.2.

(a) Find κ_F to three decimals, the charge a 1.00 µm standard wafer needs, and its bulk step at 14 A. *Checkpoint:* 0.534 A h per µm; 0.545 A h; 125.2 s.

(b) A thick-copper wafer of 2.00 µm runs its bulk step at 28 A. Find its bulk step, and the thickness it gets in case 1 if the manual bulk step stays at the standard wafer's unrounded value of part (a). By what percentage is it thin? *Checkpoint:* 132.7 s; 1.893 µm; 5.35 %.

(c) One wafer leaves every 216 s. Find the cycle-average current for standard and for thick-copper wafers. *Checkpoint:* 9.09 A and 18.17 A.

10.5 Theory and calculation · 8 min

Take the standard load of 9.09 A, $V = 200\,\mathrm{L}$, the accelerator at its make-up strength $S_V = 6.0\,\mathrm{mL/L}$, $c_S = 5.0\,\mathrm{mL}$ per A h and $d_{S,25} = 1.0\,\%/\mathrm{h}$ at 25 °C.

(a) From Equation (10.1), find the accelerator consumed at the electrodes and broken down in the bulk, in mL/h. Why do the bleed and its virgin make-up drop out at this state? Find the dose that holds the bath and the turnover time of the 1200 mL inventory. *Checkpoint:* 45.4 mL/h and 12.0 mL/h; 57.4 mL/h; 20.9 h.

(b) The feedforward uses $\hat{c}_S = 4.75\,\mathrm{mL}$ per A h. Find F_S from Equation (10.8) and the trim that makes up the rest. *Checkpoint:* 55.16 mL/h; 2.27 mL/h.

(c) From Equation (10.5), find the window's half-width as a percentage of the ratio, $e^{w(B)} - 1$, at $B = 22$ and 27 mL/L. After the overdose of case 8 the leveler is 4.5 mL/L. Find the new center r_c and the lower edge at $B = 22\,\text{mL/L}$, to two decimals, and say where a bath held at a ratio of 3.0 sits. *Checkpoint:* 9.8 % and 3.6 %; 3.67 and 3.35; below the window.

10.6 Theory and calculation · 8 min

Use the report-to-report model of Section 10.6 at the operating point of case 3 after the new lot: $c_S = 6.25\,\text{mL}$ per A h, $I = 9.09\,\text{A}$, $S_V = 6.0\,\text{mL/L}$, a breakdown of 1.0 %/h, a bleed of 2.28 L/h, a drag-out of 0.5 L/h, $V = 200\,\text{L}$, $T = 2\,\text{h}$ and $\theta = 42\,\text{min}$.

(a) Find k in L/h and the time constant $\tau = V/k$. *Checkpoint:* 14.24 L/h; 14.0 h.

(b) Find a, β_0 and β_1 of Equation (10.11), with the β in h/L to three significant figures, and check that $\beta_0 + \beta_1 = (1-a)/k$. *Checkpoint:* $a = 0.867$; $\beta_0 = 6.21 \times 10^{-3}\,\text{h/L}$, $\beta_1 = 3.11 \times 10^{-3}\,\text{h/L}$; both sides 9.32×10^{-3} h/L.

(c) Find the share $\beta_1/(\beta_0 + \beta_1)$ for $\theta = 42\,\text{min}$ and for $\theta = 60\,\text{min}$, and use it to explain why the limit falls from 210 to 141 as the age grows from 42 min to an hour. *Checkpoint:* 0.334 and 0.482.

10.7 App challenge · 8 min

Open scenario 4, *4 · A lot mix change*, at the default seed, and run each day at a speed of 60 min/s (24 s a day). Read the scorecard tiles and the exported CSV, whose time column is `Time (h)`.

(a) Run the full 24 h as shipped. Read the Voids and Worst fill margin tiles, and the accelerator dose `dose_S_mLh` in the CSV rows at 4.0 and 6.0 h, within 0.3 mL/h. *Checkpoint:* 0 voids, 0.23; 57.4 mL/h and 101.8 mL/h.

(b) Press Reset, set Dosing strategy to FB only on the DOSE-100 card before pressing Run, and run 24 h. Read the two tiles, the CVS reading `S_cvs_mLL` in any row from 4.7 to 6.6 h and from 6.7 to 8.6 h, within 0.01 mL/L, and the first row in which `metric_voids` exceeds zero, within 0.05 h. *Checkpoint:* 97 voids, −0.06; 5.96 mL/L and 5.70 mL/L; 8.1 h.

(c) Explain from the sample times why the dose on feedback only first responds to the lot at 6:40 (the step of under 1 mL/h at 4:40 answers a result from before the lot), and why the voids continue for hours after that. *Checkpoint:* The 4:40 result describes the bath of 3:58; the trim then adds about 10 mL/h a report against the roughly 45 mL/h the lot needs.

10.8 App challenge · 8 min

Open scenario 3, *3 · CVS trim on a sampled, delayed analysis*, at the default seed, and run each day for 24 h at 60 min/s.

(a) Run it as shipped. Then press Reset, set Accelerator trim gain on the AIC-S card to 0 before pressing Run, and run again, and repeat with 300 and with 400. Record the Voids and Worst fill margin tiles of the four days. *Checkpoint:* 0 and 0.13; 163 and −0.09; 16 and −0.04; 37 and −0.21.

(b) In the CSV of the day at 400, find the highest and the lowest true accelerator `S_true_mLL` after 12 h and their times, within 0.02 mL/L and 0.1 h. *Checkpoint:* 6.53 mL/L at 16.7 h; 5.12 mL/L at 22.7 h.

(c) Read ρ at gains of 300 and 400 from the solid curve of Figure 10.3, and explain why the larger gain voids more wafers although it corrects harder. *Checkpoint:* About 1.07 and 1.13, both above one.

10.9 App challenge · 8 min

Open the challenge, *Challenge · a day of 400 wafers*, at the default seed, and run each day for 24 h at 60 min/s. Use Equation (10.6) with the 1868 mL of breakdown products the day forms under every bleed policy.

(a) Run it as shipped and read the Voids, Worst fill margin, Virgin make-up and Day target tiles. *Checkpoint:* 226, −0.37, 12.0 L, not met.

(b) Press Reset. Before Run, set Dosing strategy to FF + trim (the Outlier gate stays On), Bleed policy to Hold B and Breakdown products setpoint to 24 mL/L. Run the day and read the same tiles and the true breakdown products `B_true_mLL` in the last CSV row. Compute the mean concentration at which the breakdown products left, to one decimal. *Checkpoint:* 0, 0.12, 57.0 L, met; 24.55 mL/L; 23.8 mL/L.

(c) Repeat with the setpoint at 26 mL/L, and compute the make-up saved per mL/L of setpoint between 24 and 26. *Checkpoint:* 0, 0.05, 38.7 L, met; 26.48 mL/L, 25.1 mL/L; about 9.2 L per mL/L.

10.10 Application area · 4 min

A process engineer must choose the breakdown-product setpoint for the day of the challenge, with virgin make-up at $20 per liter.

(a) From Table 10.6, price a day of make-up at setpoints of 22, 24 and 26 mL/L. *Checkpoint:* $1664, $1140 and $774.

(b) Which setpoint would you run, and what does the seed spread of Section 10.7 add to the decision? *Checkpoint:* 24: across ten other seeds its worst margin stays between 0.083 and 0.176.

10.11 Application area · 3 min

In case 8 a leveler pump runs away and puts 300 mL into the bath.

(a) Why can the leveler loop not undo the overdose, and how does the leveler leave the bath? *Checkpoint:* A pump can only add; the leveler leaves by consumption, breakdown and bath taken out.

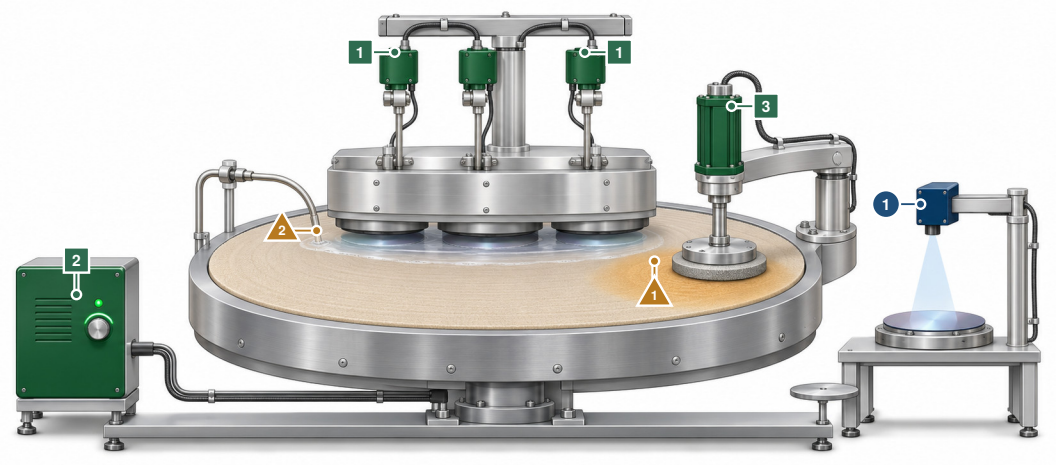
(b) Compare the two recoveries of Section 10.7 in voids, wafers plated and make-up, and name one safeguard that would have prevented the event. *Checkpoint:* Bleed 4 L/h: 197 voids, 400 wafers, 100.7 L; dump: 2 voids, 366 wafers, 216.5 L.

10.12 Review the agent · 5 min

An AI assistant reviews scenario 3 and writes: *"The trim decays by only 0.900 per analysis at its gain of 30, so it takes most of a day to correct the new accelerator lot. The CVS repeats to 2 %, which is precise enough to act on hard. Raise the trim gain to 300: the loop will then correct the lot within one or two analyses, and a thick-copper lot will be caught the same way."* Find the flaw, and say how you would check the claim.

CHAPTER 11

A Chemical Mechanical Planarization Simulator for Profile Control Across a Consumable Lifecycle



The controller moves 1 air pressures in the eight rings of the carrier head 2 the polish time of each wafer 3 how hard the diamond disc conditions the pad

The controller reads 1 a gauge that maps the polished film's thickness, one or two lots later

What disturbs it 1 the pad wearing out, and a new pad that behaves differently 2 a change in how active the slurry is

Notice that the polished film is measured only by the gauge at the right, and its map returns one or two lots after the wafer left the pad. Illustration, not to scale.

IN PLAIN TERMS **Polishing Flat on a Pad That Wears Out**

The main idea. A polisher rubs a thin film on a silicon disk down to an even thickness with a spinning pad and a gritty liquid. A head with rings of air presses harder where more film must come off.

The challenge. The pad wears with every disk and is replaced every few days by one that works a bit differently. The controller must decide what it still knows about the machine and what it must learn again.

Why it matters. Learning a new pad from one test disk, not a fixed count of six, saves silicon at every change.

THREE WAYS TO READ THIS CHAPTER

Explore	12 min	The first page, the fact boxes of the Introduction and of Section 11.6, the history boxes of Section 11.2, the try-it boxes of Section 11.6, and the key points
Study	50 min	The whole chapter with its pause points and the twelve exercises
Practice	20 min	The first page, Section 11.6, Section 11.7, Section 11.8 with its open challenge

The control question Eight carrier head chambers drive thirty-three radial thicknesses while the pad doing the work wears out and is replaced every few days. Which estimator states should survive a pad change, and how many qualification wafers does the answer cost?

Learning objectives

- Explain why eight concentric chambers cannot reach every radial profile, and separate a profile error into its reachable, unreachable and azimuthal parts.
- Compute the relative pad speed across the wafer, the settled pad activity under a conditioning dose and the removal lost to structural wear.
- Distinguish what feedforward from an incoming map and run-to-run feedback on the removal can each correct, and choose the EWMA level and trend weights against a metrology queue.
- Compare soft, carry-all and cold resets at a pad change, and compute how much one mapped wafer says about the gain of a fresh pad.
- Judge a release on predicted probability against a fixed qualification count, and say when a surrogate measurement may be trusted.

Prerequisites Sections 1.4.3, 1.4.4, 1.4.5 and 1.4.7 of Chapter 1, and Section 1.6.4 for the scenarios; Chapter 6 for EWMA recipe control through a queue.

The app https://apmonitor.com/apps/cmp-polish/

See also Volume 8, *Batch, Run-to-Run, and Hybrid Control*: Volume 8, Chapter 9 (a run-to-run estimate that knows which command each late result answers) and Volume 8, Chapter 11 (a resin capacity that cleaning restores only in part). Volume 6, *Multivariable and Plantwide Control*: Volume 6, Chapter 6 (a regularized inverse of an ill-conditioned response). Volume 5, *Delays, Slow Measurements, and State Estimation*: Volume 5, Chapter 10 (a soft sensor whose bias a delayed laboratory corrects).



11.1 Introduction

Most control problems assume that the plant is still the same plant tomorrow. A polisher is not. The pad that removes the film is a consumable, it is conditioned back and worn down on every wafer, and after a few thousand wafers somebody takes it off the machine and puts a new one on. The tool that comes back is the same carrier head with the same eight pneumatic chambers, the same footprints, the same metrology and the same product, and it is a different process. What the control system learned about the old pad is partly still true and partly worthless, and the boundary between those two parts is where this application lives.

> ✦ DID YOU KNOW? **A Billion-Dollar Consumable**
>
> In 2005 the slurries, pads and pad conditioning materials that polishers use up were a market worth 1.1 billion US dollars, expected to grow to 1.9 billion by 2009 [1]. The pad of this chapter is one of those consumables, and it is replaced every few days.

The obvious answer is the one the industry mostly uses. Run a fixed number of sacrificial qualification wafers after every pad change, throw them away, and resume. That is safe, it is simple, and it costs scrapped silicon at every maintenance event on every polisher in the fab. The alternative is to say out loud which estimator states belong to the part that was replaced and which belong to the machine, reset only the first group, widen the covariance on what was reset, and then release production on evidence rather than on a count. The three obvious ways to handle a pad change differ by more than an order of magnitude in what they cost, and Section 11.6 measures all three on identical hidden pads.

The same shape appears outside semiconductor manufacturing. A catalyst is replaced, a membrane module is changed, an analyzer is recalibrated, a mill liner is renewed, and in every case a control system that has been adapting for months is asked what it still believes. The literature on run-to-run control is largely written about drift, where the estimator tracks a slowly moving target [2, 3], as

Section 1.4.3 sets out and Chapter 6 measures on a patterning line. A maintenance event is not drift. It is a known instant with an unknown outcome, and the correct response to it is a reset of some states and an increase of uncertainty on others rather than a change of gain.

Underneath that sits a second problem that a student meets first. Eight chambers drive thirty-three radial calculation points. Nobody gets to choose the thickness at each radius independently, the response to one chamber covers a broad band of radii, and a profile error that will not come out is often an unreachable spatial mode rather than evidence of poor tuning [4]. Equal chamber pressure does not even give a flat wafer, because relative pad speed rises toward the wafer edge and Preston behavior is linear in velocity [5, 6].

This chapter makes the cost of a maintenance event measurable on an interactive polisher. On it a retained population prior and one mapped monitor wafer release production where a fixed policy needs six, and the same single wafer applied through the wrong reset strategy puts the first production lot 41.7 nm off target. The simulator is an eight-state consumable model driving a thirty-three by eight spatial plant, with a constrained recipe optimizer, an age-aware residual estimator and a Bayesian fresh-pad update written closely enough that a reader can rebuild them. The analysis decomposes a profile error into its reachable and unreachable parts, and it shows that the information in a qualification wafer comes from the retained prior and the thirty-three sites rather than from the excitation design. On one plant, one seed and one set of hidden pad parameters, three reset strategies that are all defensible in isolation produce 2.5 nm, 41.7 nm and 59.4 nm of first-lot error. Preston's relation, the exponentially weighted moving average and Bayesian parameter estimation are all standard methods.

Section 11.2 places the problem in its domain. Section 11.3 gives the simulation model: the influence matrix of the carrier head, the removal law, the two pad states and the metrology. Section 11.4 describes the controllers as implemented, from the recipe solve to the qualification supervisor. Section 11.5 analyzes what a pad change may delete, and Section 11.6 reports the measured outcomes of the scenarios. Section 11.7 describes the hardware such a polisher runs on, Section 11.8 states the limitations and Section 11.9 concludes.

11.2 From Pad Wear to Consumable-Aware Recipe Control

Polishing as a controlled process is older than the semiconductor industry that adopted it. Its removal law came from plate glass and remains the first-order description of what a polisher does [5], with exponents near but not exactly unity [6] and the chemistry added by later models [7], as the box below tells.

> ⧖ HISTORY **From Plate Glass to Wafers (1927 to 2007)**
>
> In 1927 a study of the theory and design of plate glass polishing machines related the rate at which glass is removed to the pressure applied and to the relative velocity between the pad and the glass. Most models of chemical mechanical planarization are still inspired by that equation, although it is empirical and the effect of the abrasive's hardness and size on its coefficient is unclear [1]. A modified model of 1997 predicted that removal grows with pressure to the power five sixths and with speed to the power one half, and experiments on thermal oxide agreed [6]. By 2007 a review called the process the only technique able to give excellent local as well as global planarization [1].

Feedback on the recipe arrived in the middle 1990s, and it arrived because of the pad. Run-by-run compensation for pad wear was demonstrated on a five-hundred wafer experiment in 1996 [8], on statistical machinery developed just before [2] and analyzed just after [3], and within five years the estimator had been given the pad's age [9]. The practice of that decade is collected in a single volume [10], and the box below tells the story.

> ⧖ HISTORY **Run-by-Run Control Comes to the Polisher (1995 to 2001)**
>
> In 1995 a run-by-run controller combined statistical process control with feedback. Its rapid mode adapted to sudden shifts such as those caused by maintenance, and its gradual mode to slow drifts [2]. In 1996 a prototype system controlled a single-wafer polisher lot by lot. It measured film thickness at nine points on blanket oxide wafers of 8 inches, adapted its model with an exponentially weighted moving average, and generated recipes within bounds on the inputs and the steps of the machine settings. It compensated a substantial drift in the tool's removal rate, and its model held on more than one tool for several months [8]. By 2001 an age-based double EWMA was applied to polishing, because an EWMA alone does not suffice for a process that wears out [9].

Profile control is younger. A carrier head with several independently pressurized concentric chambers turns one manipulated variable into a small vector, and the plant it drives is tall because the film is measured at many radii. Multizone polishing has been treated as exactly that, many radial outputs driven by fewer pressures through local Preston coefficients, with the inverse taken through a pseudoinverse or a singular value decomposition [4]. The mechanical question of how chamber pressure becomes contact pressure at the film has its own literature [11, 12], and the pad surface that the pressure acts through is itself a controlled variable maintained by diamond disc conditioning [13].

The newest layer is metrology that is not metrology. Spatially resolved removal and non-uniformity can be predicted from equipment traces and consumable counters on production data sets [14], and virtual metrology has been used to close run-to-run loops on wafers that were never measured [15]. The accompanying risk has also been named. A surrogate that feeds its own corrections is a closed loop with no ground truth in it, and the practical answer has been to let the confidence of the prediction decide the physical sampling rate rather than to let the prediction replace the sample [16].

What none of that literature makes easy to see is the moment of replacement. The estimator papers treat the pad as an aging parameter, the profile papers treat the head as a fixed matrix, and the virtual metrology papers treat the tool as stationary enough to train on. The application of this chapter puts all three on one machine and then changes the pad.

11.3 Process Model

A 300 mm wafer is held face down by carrier head CH-101 against a rotating polyurethane pad on platen PL-101. Slurry is delivered to the pad surface, diamond disc conditioner CD-101 sweeps the pad to restore its texture, and the carrier head presses the wafer through eight concentric membrane chambers at pressures p_1 through p_8. The wafer is polished for a common time T_p, unloaded, and replaced by the next one. The equipment and every symbol that appears in the equations are drawn in Figure 11.1.

Table 11.1 collects what the controller acts on and what it acts against. The product factor η_{pr} is the one disturbance the controller knows, because the product identity is on the lot record. Everything else in the lower half of the table is unmeasured, and the only evidence that any of it moved is a thickness map arriving one or two lots later. Time is carried in minutes, thickness in nanometers, pressure in pounds per square inch and polish time in seconds. Eight states are integrated continuously and the per wafer arithmetic is evaluated once every $T_{takt} = (T_p + t_{ovh})/60$ minutes, which is 1.5 min at the design recipe.

11.3.1 Geometry and the Influence Matrix

The wafer is represented by $M = 33$ radial calculation points on a midpoint grid, $\rho_m = (m - 0.5)/M$ for $m = 1 \ldots M$ with $r_m = \rho_m R$ and $R = 150\,\text{mm}$. Area weights $w_m \propto \rho_m$ are normalized to sum to one, so every area-weighted mean in this chapter is an honest wafer average rather than a radial average.

Eight chambers occupy normalized radial bands with edges at 0, 0.300, 0.500, 0.645, 0.765, 0.860, 0.925, 0.968 and 1. Interior chambers are wide and edge chambers are narrow, which is how

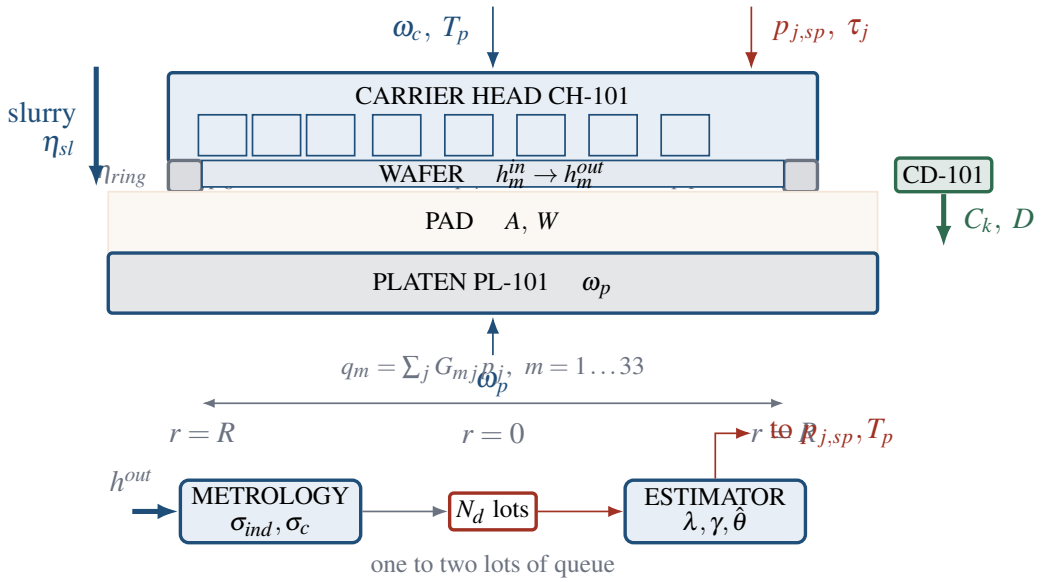


Figure 11.1 The polisher and its measurement path. Eight concentric carrier head chambers at pressures p_1 through p_8 press the wafer against the pad through an influence matrix G that spreads each chamber over a band of radii. The pad carries two states that are not the same thing, an active surface A that conditioning restores and a structural wear W that nothing restores. The post-polish thickness map is sampled and returns one to two lots late.

Table 11.1 Controlled, manipulated and disturbance variables

Symbol	Role	Design value	Range
$\bar{h}^{out}$	CV, N_d lots late	700.00 nm	400–1100 nm
h_1^{out}	CV, wafer center	703.59 nm	—
h_{33}^{out}	CV, wafer edge	696.64 nm	—
T_p	MV, once per wafer	60.0 s	35–90 s
$p_1 \ldots p_8$	MV, once per wafer	3.60 psi	2.5–4.8 psi
C_k	MV, once per wafer	0.55	0–1.5
A	state, unmeasured	0.9083	0–1
W	state, unmeasured	0 at install	0–1
D	state, unmeasured	0 at install	0–1
E	state, counter	0 at install	wafer-equivalents
η_{sl}	state, DV, unmeasured	1.000	0.80–1.20
η_{pr}	state, DV, known	1.000	0.75–1.25
η_{ring}	state, DV, unmeasured	1.000	0.70–1.30
β_M	state, DV, unmeasured	0 nm	±15 nm
θ	hidden, per pad	**0** (mean)	$\sigma_{\theta g}$, $\sigma_{\theta s}$
h^{in}	DV, measured pre-polish	1200 nm	±80 nm

multizone heads are built. Each column of the influence matrix is a Gaussian footprint centered on its chamber,

$$\tilde{G}_{mj} = \exp\left[-\frac{(\rho_m - c_j)^2}{2\sigma_j^2}\right], \qquad \sigma_j = 0.9\,\frac{b_{j+1} - b_j}{2} + 0.045 \tag{11.1}$$

where c_j is the midpoint of band j and b_j are the band edges. The matrix is then row-scaled,

$$G_{mj} = \frac{\tilde{G}_{mj}}{\sum_{\ell} \tilde{G}_{m\ell}}, \qquad q_m = \sum_{j=1}^{8} G_{mj}\, p_j \tag{11.2}$$

so that eight equal pressures reproduce a uniform contact field exactly. That scaling is a calibration of the whole matrix and not a per-column normalization, and the columns keep their different shapes and their different totals. A 0.20 psi step on interior chamber 4 raises the contact pressure by 0.095 psi at its own center, stays above half of that across 21 % of the wafer radius (the span of the calculation points above half height), and is still 38 % and 56 % of its peak at the centers of chambers 3 and 5. The same step on edge chamber 8 peaks at 0.081 psi and stays above half of it across 3 % of the radius. The edge chamber has localized authority because it is narrow rather than because it is strong. The bake plate of Section 1.4.5 has the same structure in heat, with a conductance matrix in place of G.

Pause. Try Exercise 11.3 before reading on.

11.3.2 Kinematics and Removal

The platen turns at ω_p and the carrier at ω_c about centers a distance d apart, so the azimuthally averaged relative speed at radius r_m is Equation (11.3).

$$v_m = \sqrt{(\omega_p d)^2 + (\omega_p - \omega_c)^2\, r_m^2} \tag{11.3}$$

At 80 rpm, 60 rpm and $d = 0.220$ m this is 1.8431 m/s at the innermost calculation point and 1.8689 m/s at the outermost, a ratio of 1.0140. The reference speed v_{ref} is the area-weighted mean of v_m at the nominal recipe.

Local removal follows a spatial Preston relation with multiplicative state corrections, Equation (11.4).

$$RR_m = RR_{ref}\, \eta_{pad,m}\, \eta_{sl}\, \eta_{pr}\, \eta_{ring,m} \left(\frac{v_m}{v_{ref}}\right)^{\beta} \left(\frac{q_m}{p_{ref}}\right)^{\alpha} \tag{11.4}$$

Integrated removal over a polish of T_p seconds is $r_m = RR_m T_p/60$ in nanometers, and the outgoing thickness is $h_m^{out} = h_m^{in} - r_m + \varepsilon_{proc}$ where ε_{proc} is a wafer-to-wafer process deviate of standard deviation σ_{proc} common to all radii. The retaining ring acts only at the outer edge, $\eta_{ring,m} = 1 + (\eta_{ring} - 1)\psi_e(\rho_m)$ with $\psi_e(\rho) = \exp[-((1-\rho)/0.08)^2]$.

The pressure ramp is paid for rather than discarded. Each chamber is a first order actuator of time constant τ_j that falls linearly from $\tau_{p,max}$ at chamber 1 to $\tau_{p,min}$ at chamber 8, and the time-averaged pressure over one polish has the closed form of Equation (11.5).

$$\bar{p}_j = p_{j,sp} - \frac{\tau_j}{T_p}\,(p_{j,sp} - p_{j,prev})\left(1 - e^{-T_p/\tau_j}\right) \tag{11.5}$$

Equation (11.5) is exact for the integrated removal when the pressure exponent α is unity, and is a first order approximation away from it. The contact field of Equation (11.2) is evaluated at $\bar{p}_j$.

11.3.3 The Pad Is Two States

The pad carries an active surface A that conditioning restores and a structural wear W that nothing restores. Conditioning is delivered in situ at dose C_k per wafer and mechanical exposure accumulates with pressure, speed and time, so both are written as rates per minute using ΔE_k from Equation (11.6).

$$\Delta E_k = \frac{T_p}{T_{ref}} \frac{\bar{v}_k}{v_{ref}} \frac{\bar{q}_k}{p_{ref}}, \qquad \dot{E} = \frac{\Delta E_k}{T_{takt}} \tag{11.6}$$

The three consumable states then obey Equations (11.7) through (11.9), in which $\dot{C} = C_k / T_{takt}$ is the conditioning dose rate.

$$\dot{A} = k_c \, \eta_{cond} \, \dot{C} \, (1 - A) - k_g \, \dot{E} \, A \tag{11.7}$$

$$\dot{W} = \frac{\dot{E} + \rho_c \dot{C}}{N_{EOL}}, \qquad N_{EOL} = N_{pad} \, (1 + \rho_c C_{nom}) \tag{11.8}$$

$$\dot{D} = \frac{\dot{C}}{N_{cond}}, \qquad \eta_{cond} = \max[\eta_{c,min}, \, 1 - (1 - \eta_{c,min}) D] \tag{11.9}$$

Equation (11.7) settles at $A^* = k_c C_{nom} / (k_c C_{nom} + k_g) = 0.9083$ under the default dose, which is the value the reference removal rate is calibrated at. The capacity N_{EOL} in Equation (11.8) is scaled so that the quoted pad life is the life obtained under the default conditioning policy rather than a polishing-only count, which matters because conditioning consumes the pad. The affinity resin of Volume 8, Chapter 11 has the same split, a capacity that cleaning restores only in part.

The pad activity multiplier assembles the two states with the hidden pad-lot correction, Equation (11.10).

$$\begin{aligned} \eta_{pad,m} = e^{\theta_g} \, f_A(A) \, \left(1 - d_w W^{1.4}\right) \\ \times \left[1 + \theta_t \phi_{t,m} + \theta_c \phi_{c,m} + \theta_e \phi_{e,m}\right] \end{aligned} \tag{11.10}$$

with $f_A(A) = (0.80 + 0.25A) / (0.80 + 0.25A^*)$, which is unity at the running operating point rather than at an idealized $A = 1$. The exponent 1.4 on structural wear is why the decline accelerates. On a 260 wafer pad a fixed recipe holds the mean inside the ± 14 nm band until wafer 70, the first to leave it, and then finishes 90 nm above target at end of life.

11.3.4 Profile Modes and the Hidden Pad Parameters

Three shape functions carry the pad-to-pad variation. Each is centered by subtracting its own area-weighted mean on this grid, which is below 0.001 for the two polynomials and 0.167 for the edge bump, so all three have exactly zero area-weighted mean. They move the profile without moving the mean, which is what allows the mean and the shape to be released against separate chance constraints later.

$$\begin{aligned} \phi_t = 2\rho^2 - 1 - \bar{\phi}_t, \qquad \phi_c = 6\rho^4 - 6\rho^2 + 1 - \bar{\phi}_c, \\ \phi_e = e^{-\left(\frac{1-\rho}{0.10}\right)^2} - \bar{\phi}_e \end{aligned} \tag{11.11}$$

Together with the constant they are orthonormalized under the area weights at load time, giving a basis $\{\mathbf{b}_0 \ldots \mathbf{b}_3\}$ on which a measured residual is projected by a weighted dot product rather than by a least squares solve. At every pad replacement the hidden vector $\theta = [\theta_g, \theta_t, \theta_c, \theta_e]^T$ is drawn once from the population, with standard deviations $\sigma_{\theta g}$ for the gain and $\sigma_{\theta s}$, $0.8\sigma_{\theta s}$ and $1.6\sigma_{\theta s}$ for tilt, curvature and edge, and held fixed for that pad.

11.3.5 Quality, Metrology and Disturbances

Two non-uniformity conventions are reported because the literature quotes both, and both are computed on the outgoing thickness rather than on the removal amount.

$$NU_\sigma = 100 \, \frac{\sigma_h}{\bar{h}^{out}}, \qquad WIWNU = 100 \, \frac{h^{out}_{max} - h^{out}_{min}}{2 \, \bar{h}^{out}} \tag{11.12}$$

Blanket-monitor practice quotes removal-rate non-uniformity, and for an unpatterned wafer with a flat incoming film the two agree. They do not agree here, because correcting an incoming profile requires deliberately non-uniform removal and a removal-based figure would penalize the entire purpose of the eight chambers.

A physical map carries an independent per-site deviate of standard deviation σ_{ind} and a correlated whole-map deviate of standard deviation σ_c, so its covariance is $\mathbf{R} = \sigma_{ind}^2 \mathbf{I} + \sigma_c^2 \mathbf{1}\mathbf{1}^T$ plus the process

Table 11.2 Model parameters. All are adjustable in the model dialog unless marked hidden, which means the interface does not expose them, or fixed, which marks the geometry.

Symbol	Meaning	Value	Unit
R	Wafer radius (fixed)	150	mm
M	Radial calculation points (fixed)	33	—
N_z	Carrier head chambers (fixed)	8	—
N_{lot}	Wafers per lot (hidden)	25	wafer
t_{ovh}	Load, unload and index	30	s
N_{pad}	Nominal pad life	1500	wafer
RR_{ref}	Reference mean removal rate	500	nm/min
p_{ref}	Reference chamber pressure	3.60	psi
α	Pressure exponent	1.00	—
β	Velocity exponent	1.00	—
h_0^{in}	Nominal incoming thickness	1200	nm
h^{target}	Target final thickness	700	nm
h_{min}	Minimum remaining film	640	nm
$\tau_{p,min}$	Fastest chamber time constant	0.5	s
$\tau_{p,max}$	Slowest chamber time constant	3.0	s
ω_p	Platen speed	80	rpm
ω_c	Carrier speed	60	rpm
d	Carrier offset	0.220	m
k_c	Conditioning recovery	0.18	/dose
k_g	Glazing coefficient	0.010	/wafer-eq
d_w	End-of-life removal loss	0.18	—
ρ_c	Pad consumed per dose	0.25	wafer-eq/dose
N_{cond}	Conditioner disc life	2500	dose
$\eta_{c,min}$	Conditioner efficiency floor	0.50	—
C_{nom}	Dose the pad life is quoted at	0.55	—
t_{pre}	Required break-in conditioning	8	min
$\sigma_{\theta g}$	Fresh-pad gain spread	0.05	1σ
$\sigma_{\theta s}$	Fresh-pad shape spread	0.025	1σ
σ_{ind}	Independent site noise	2.0	nm
σ_c	Correlated map uncertainty	3.0	nm
σ_{VM}	Virtual metrology uncertainty	5.0	nm
σ_{proc}	Wafer-to-wafer process noise	1.2	nm
σ_{in}	Incoming variation	6.0	nm
τ_{dist}	Disturbance relaxation (hidden)	1.0	min
e_h^{max}	Mean thickness tolerance	14	nm
$RMSE^{max}$	Profile error limit	12	nm
$WIWNU^{max}$	Range non-uniformity limit	1.8	%

term. Maps are taken on the first wafer of each lot and on every $\lfloor N_{lot}/n_s \rfloor$-th wafer after it, up to n_s maps per lot, so two maps per lot are wafers 1 and 13. Maps arrive N_d lots later, except during qualification and the first released lot, when measurement is immediate. Section 1.4.4 gives the cost of such a queue in general. Virtual metrology predicts the same map from the controller model calibrated at the last physical anchor, with its own uncertainty σ_{VM} and a switchable bias.

Four disturbances relax to their slider values on a time constant τ_{dist}, which keeps a step continuous enough for a fixed-step integrator. The incoming profile is $h_m^{in} = h_0^{in} + \delta_0 + a_t\phi_t + a_c\phi_c + \varepsilon_{in}$ with a_t scaled to 40 nm at full slider travel and a_c to 30 nm. The azimuthal disturbance is separate. It has zero radial mean by construction, contributes its own root-mean-square of $a_{az}/\sqrt{2}$ in quadrature with whatever the radial loop achieves, and no concentric chamber can touch it.

11.3.6 Initialization and Numerics

The eight states initialize to $A = A^* = 0.9083$, $W = 0$, $D = 0$, $E = 0$, $\eta_{sl} = \eta_{pr} = \eta_{ring} = 1$ and $\beta_M = 0$, which is a freshly qualified pad. A scenario may pin any of them, and the pad replacement action sets W, E to zero and A to 0.15, because a new pad is smooth rather than conditioned. Integration is fixed-step fourth order Runge-Kutta at $\Delta t = 0.05$ min with the controller on the same scan and the historian logging twice per minute, so every trend and every exported record reproduces exactly on any machine (Section 1.6.1). Matrix work is done with Cholesky factorizations and solves, including the four by four posterior covariance, which is formed column by column from the factor. The eight by eight recipe solve and the four by four estimator update both run once per wafer, which is once every 1.5 min of simulated time at the design recipe.

11.3.7 Assumptions

The head is a fitted influence matrix with Gaussian footprints and not a finite element model of membrane, wafer, pad and retaining ring. A real identified matrix will not be as smooth as this one, and its columns will not be symmetric about their chamber centers.

Preston behavior is the nominal relationship and the exponents default to unity. The pressure dose of Equation (11.5) is exact only at unit pressure exponent, so away from $\alpha = 1$ the integrated removal carries a first-order approximation of the ramp.

Pad activity, structural wear and conditioner efficiency are low-order latent states chosen to reproduce break-in, glazing, slow decline and a pad-to-pad reset. Their coefficients are not fitted to any named pad, slurry, conditioner or tool, and the quoted pad life is a configurable teaching value and not a prediction.

The consumable states integrate continuously at the per wafer average rate rather than stepping once per wafer. For in situ conditioning that is the physics. For ex situ conditioning between wafers it smooths a sawtooth the application never shows.

The wafer is thirty-three radial points. Azimuthal and die-scale variation enter only as an uncorrectable contribution to the graded profile error, so dishing, erosion, defects, scratches and electrical yield are all outside the model.

The fresh-pad prior assumes a repeated pad family with stable metadata. It should be widened or abandoned after a material, supplier, conditioner, head or recipe change, and the application offers no mechanism for detecting that such a change has happened.

11.4 Controllers as Implemented

Four layers run on four clocks. Chamber pressure settles in seconds inside one wafer, the recipe is written once per wafer, the run-to-run residual updates when a measurement arrives one or two lots later, and the fresh-pad estimator updates once per pad. The recipe authority control selects which of the middle two layers is active, and the header MANUAL and AUTO switch selects whether any of them is.

11.4.1 What the Controller Believes

The controller reads the pad counters and the product identity and nothing else. Its model state is A and W from the counters, η_{pr} from the lot record, unity for η_{sl} and η_{ring}, and $\hat{\theta}$ from the estimator. The gap between that and the true state of Table 11.1 is exactly what the run-to-run residual has to absorb, and naming it explicitly is what makes the residual interpretable.

11.4.2 The Constrained Recipe

With $\alpha = 1$ the predicted removal is linear in the pressure vector, so the recipe solve is a small regularized least squares problem. Writing a_m for the removal per unit contact pressure per sixty seconds at the controller's model state and $\hat{\mathbf{e}}$ for the age residual of Section 11.4.4, the target is

$\mathbf{t} = \mathbf{h}^{in,est} - h^{target}\mathbf{1} - \hat{\mathbf{e}}$ and the solve is Equation (11.13).

$$\min_{\mathbf{p}} \quad \sum_m w_m \left(a_m \frac{T_p}{60} \sum_j G_{mj} p_j - t_m \right)^2 + \lambda_p \|\mathbf{p} - p_{ref}\mathbf{1}\|^2 + \lambda_\Delta \|\mathbf{p} - \mathbf{p}_{prev}\|^2 + \lambda_s \|\mathbf{D}_2 \mathbf{p}\|^2 \tag{11.13}$$

where $\mathbf{D}_2$ is the second difference operator on the eight chamber pressures and $\lambda_p = 0.05\lambda_s$, $\lambda_\Delta = 0.10\lambda_s$. The smoothness weight is what prevents an ill-conditioned inverse from answering a high-frequency radial target with a sawtooth, the problem that Volume 6, Chapter 6 meets at the scale of a ninety-six by ninety-six response. Its scale matters more than it looks. The diagonal of the data term of the normal-equation matrix runs from about 360 for the edge chamber to about 1280 for chamber 2 at the default removal rate and a 60 s polish, so a second difference weight in the hundreds does not smooth the recipe, it replaces it, and the optimizer then returns eight equal pressures for every target it is given. The shipped default is $\lambda_s = 8$.

Polish time carries the level and the chambers carry the shape, so the solve alternates. Four iterations of solving Equation (11.13) at the current T_p and then rescaling T_p by the ratio of desired to predicted mean removal, followed by a final pressure solve, put the predicted mean within 0.1 nm of target on 95 % of the production wafers of scenarios 3, 6 and 11. The pressure box, the neighboring-chamber gradient limit $\Delta p_{ad\,j}$ and the run-to-run move limit Δp_{move} are enforced by three sweeps of projection.

The minimum film is a chance constraint and not a nominal one. With z_α the normal quantile at the release probability and $\sigma_{h,m}$ the predicted output standard deviation of Section 11.4.5, the polish time is reduced until Equation (11.14) holds at every radius.

$$h_m^{in,est} - \hat{r}_m - z_\alpha \sigma_{h,m} \geq h_{min} \tag{11.14}$$

The film constraint is hard, so it overrides the target.

11.4.3 Feedforward

What the optimizer is allowed to know about the wafer in front of it is a control with three settings, a case of the feedforward between process steps of Section 1.4.7. The product nominal gives $h_m^{in,est} = h_0^{in}$. The lot-representative mean gives $h_m^{in,est} = h_0^{in} + \delta_0$, one number for the whole lot. The inline map gives the measured h_m^{in} at all thirty-three points. The difference between the last two is the entire value of inline pre-polish metrology, and it appears in the first wafers after a product change, as fast as the run-to-run move limit lets the recipe step, and nowhere else.

11.4.4 The Age-Aware Residual

When a measured removal map for origin wafer i arrives, the controller reconstructs the model as it stood when that wafer was polished, using the stored pad states, product factor, estimate vector, applied pressure dose and polish time. The residual $\mathbf{e}_i = \mathbf{r}_i^{meas} - \hat{\mathbf{r}}_i$ is projected onto the four orthonormal profile modes and each coefficient is updated by Equation (11.15), the EWMA of Equation (1.2) applied to each mode with a slope term in exposure added.

$$\ell_k^+ = \lambda\, c_{k,i} + (1-\lambda)\ell_k^-, \qquad s_k^+ = \gamma \frac{\ell_k^+ - \ell_k^-}{E_i - E_{prev}} + (1-\gamma) s_k^- \tag{11.15}$$

and replayed forward to the present exposure by $\hat{e}_k = \ell_k^+ + s_k^+ (E - E_i)$. Two implementation details are not optional. The slope update is skipped when the exposure interval is below one wafer-equivalent, because a rate needs a real interval to be a rate. Without that guard, and with the exposure recorded before rather than after the wafer, the first measurement of a run arrives with an exposure interval of zero and the slope goes straight to its bound. Both the level and the slope are bounded, the level at half

the reference removal and the slope at 2 nm per wafer-equivalent, which is 24 times the steepest wear slope of a default 1500 wafer pad (0.084 nm per wafer-equivalent at end of life) and still finite. On the 260 to 400 wafer pads of the scenarios the steepest wear slope is 0.32 to 0.48 nm per wafer-equivalent, four to six times below the bound.

A delayed result is evidence about the model that produced it and not about today's model, the comparison with the model of its time that Section 1.4.4 asks for. Replaying it against the controller's current product factor, pad state and estimate rather than the ones in force at the originating wafer charges whatever changed during the queue to the measurement. In the challenge scenario with the suggested settings, the first map delivered after the 10 % product change at 150 min then reports a removal residual 20 nm smaller than the true one and the second 28 nm smaller, and the loop under-corrects the earlier slurry step by that amount.

An arriving map is gated. The normalized innovation is computed against the full innovation covariance, which is the measurement covariance plus the contribution of the parameter posterior, and not against the measurement noise alone. On the qualification wafers of scenarios 7, 8 and 11 a fresh pad whose gain the controller does not yet know produces an innovation of 15 to 34 nm root-mean-square against a measurement standard deviation of 3.8 nm. Scaled on measurement noise only, that is 3.9 to 8.9 standard deviations and the gate shrinks the one wafer that was supposed to describe the pad, while scaled on the full covariance it is 0.4 to 1.2 and the wafer is used at full weight. Above three standard deviations the map is shrunk toward the prediction in Huber fashion, above nine it is rejected, and five consecutive rejections place the tool on hold, because a gate that rejects every map is an open loop wearing a safety badge.

11.4.5 The Fresh-Pad Posterior

At a qualification wafer the four fresh-pad parameters are updated in information form. With $\mathbf{H}$ the sensitivity of predicted removal to θ, whose columns are the base removal multiplied by the four basis functions, and $\mathbf{z}$ the residual against the base model with $\theta = \mathbf{0}$, the update is Equation (11.16).

$$\begin{aligned} \left(\mathbf{P}^{+}\right)^{-1} &= \left(\mathbf{P}^{-}\right)^{-1} + \mathbf{H}^{T}\mathbf{R}^{-1}\mathbf{H} \\ \mu^{+} &= \mathbf{P}^{+}\left[\left(\mathbf{P}^{-}\right)^{-1}\mu^{-} + \mathbf{H}^{T}\mathbf{R}^{-1}\mathbf{z}\right] \end{aligned} \tag{11.16}$$

The rank-one structure of $\mathbf{R}$ inverts analytically,

$$\mathbf{R}^{-1} = \frac{1}{s^{2}}\left(\mathbf{I} - \frac{\sigma_{c}^{2}}{s^{2} + M\sigma_{c}^{2}}\mathbf{1}\mathbf{1}^{T}\right), \qquad s^{2} = \sigma_{ind}^{2} + \sigma_{proc}^{2} \tag{11.17}$$

which keeps the whole update inside four by four arithmetic and guarantees that the posterior never widens on a valid measurement. That rank-one term is also the reason thirty-three radial sites on one wafer are not thirty-three wafers. It absorbs most of the information about global gain while leaving the three shape modes sharply determined.

The predicted output standard deviation used by Equation (11.14) and by the release gate separates the mean from the shape, because mode zero is a global gain and modes one to three have exactly zero area-weighted mean. Lumping them into one number makes the profile error marginal carry the gain uncertainty a second time and puts the release gate out of reach of any qualification wafer.

11.4.6 The Qualification Supervisor

The supervisor is a state machine over PAD INSTALLED, PRECONDITIONING, QUALIFICATION, RELEASED WITH ENHANCED SAMPLING, PRODUCTION, END OF LIFE and HOLD. A pad replacement holds production, zeroes W and E, sets A to 0.15, draws a new hidden θ, discards every queued result from the previous pad, and resets the production baseline recipe to the qualified nominal. What it does to the estimator is the subject of Section 11.5.

Preconditioning runs conditioning for t_{pre} minutes with no wafer consumed, which raises A from 0.15 to about 0.90 and charges the clock and the pad for it. A model-based controller can reduce the

number of qualification wafers. It cannot replace the physical work that gives a new pad a usable surface, and the application keeps the two separate on purpose.

Each qualification wafer is deliberately under-polished, at 86 % of the nominal removal demand computed against the controller's own model, and carries a balanced excitation of the three shape modes (tilt, curvature and edge) at 0.22 psi amplitude. The excitation pattern is chosen from a small safe candidate set, the nominal recipe and four sign patterns, by maximizing the log determinant of the posterior information matrix. The under-polish is not a detail. A fresh pad can be more active than the model believes (one population standard deviation of gain is 5 % of removal), so a monitor wafer run at the nominal recipe can land below target and one carrying a profile excitation on top of that can cross the minimum film limit while still being called safe. The qualification wafers of scenarios 7 and 8 keep 116 and 66 nm of film above h_{min}.

Release requires three things at once. No chamber may be faulted or saturated, the recipe must be within its limits, and the predicted joint probability of meeting the mean, profile error and non-uniformity limits must reach $P_{release}$. The joint probability is the product of three normal marginals. They do not fail independently, because a gain error moves all three together, so the number is optimistic in exactly the case where it matters most and it supplements process qualification rather than replacing it. When the qualification budget is spent without release the tool holds for engineering review, which is the correct outcome and not a failure of the policy.

Production recipes are limited to Δp_{move} per run, measured from the last production recipe rather than from the last command, because a qualification wafer is a deliberate excitation and not a recipe anybody intends to run. A pad replacement resets that baseline to the flat qualified recipe. Qualification and the first released lot carry a larger post-maintenance allowance of 0.7 psi, because the evidence for that step is a mapped wafer taken deliberately under a hold with the whole following lot measured. In scenario 7 the allowance raises the predicted pass probability at the release decision from 0.9985 to 0.9999.

11.5 What a Maintenance Event May Delete

Three strategies are offered and all three are defensible in isolation.

A *soft* reset zeroes the age residual level and slope, sets μ to the population mean, restores $\mathbf{P}$ to its prior, a standard deviation of 0.05 on each of the four parameters (the population spread of the gain, and wider than that of the three shape parameters), and keeps the influence matrix, the measurement model and the product models. It is the only one of the three under which a mapped monitor wafer has anything to update.

A *carry-all* reset keeps the worn pad's converged estimate and narrows the covariance, on the reasoning that the tool is the same tool and the estimate was hard won. The estimator is then confident about a pad that is in the bin. The release gate is computed from the controller's own belief, so it passes immediately, and the error is the difference between the two pads.

🕮 **Pause.** Try Exercise 11.2 before reading on.

A *cold* reset discards the fresh-pad model along with the pad. Nothing is left for an informative wafer to update, the posterior never contracts, and the qualification policy reverts to counting wafers. The only mechanism still running is the delayed age residual, which needs several lots to converge.

The distinction that makes this tractable is between states that belong to the part that was replaced and states that belong to the machine. The influence matrix, the neighboring-chamber coupling, the measurement covariance and the population of fresh pads survive a pad change. The wear state, the accumulated residual trend and the pad-specific correction do not. Writing that as a rule gives three instructions in order: reset what belongs to the part that was replaced, keep what belongs to the machine, and raise the uncertainty rather than the gain. Figure 11.2 sorts the estimator's states by owner and shows what each strategy does with them.

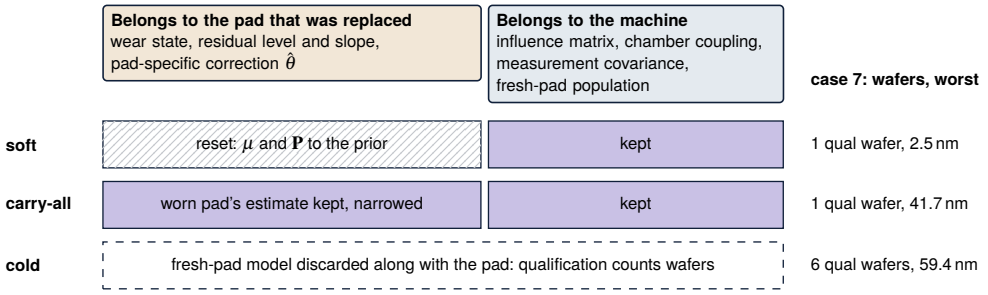


Figure 11.2 What a pad change may delete. The states of the estimator sorted by the part they belong to, and what the soft, carry-all and cold resets do with each, beside the outcome of case 7 in Table 11.3. Tinted cells are kept, hatched cells are reset, and the dashed cell is discarded. Notice that only the soft reset both clears what belongs to the pad and keeps the population that one mapped wafer can update.

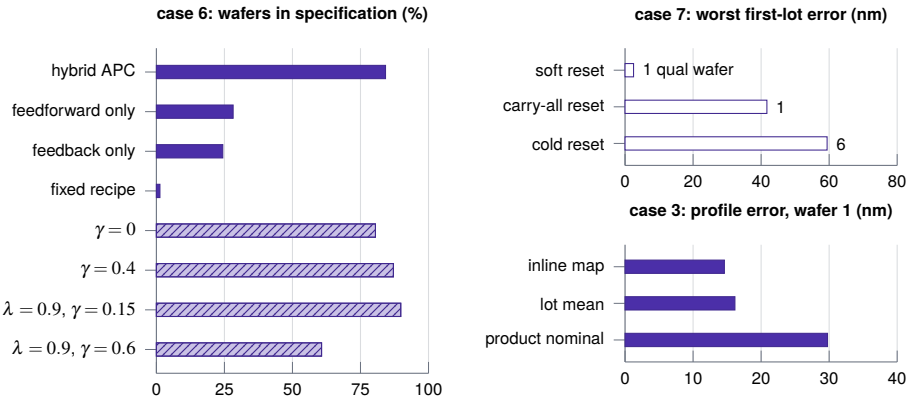


Figure 11.3 Rows of Table 11.3 drawn as bars. Left, the wafers in specification across the pad life of case 6 for the four architectures (solid bars) and for other trend and level weights of the hybrid loop (hatched bars). Right, the worst of the first twelve production wafers after the pad change of case 7, labeled with the qualification wafers spent, and the profile error of the first wafer after the product change of case 3. Notice that the two halves of the hybrid loop hold far fewer wafers alone than together, and that the cold reset spends the most wafers and lands farthest off.

The third instruction is the one that is usually skipped. A controller that resets the correct states and leaves the covariance alone is as confident after the change as before it, which is not a statement anybody can defend.

11.6 Results

Every number in this section is read from the running application at its default seed, on the shipped settings of the scenario named beside it unless a change is stated, with each change made before the run starts. A change made before the run and the same change carried through Reset give one run. Wafer counts and yields are the application's own tallies, and an event time is the instant of the controller execution at which it happens. The eleven scenarios run from manual chamber mapping to a graded lifecycle with two pad replacements, and Section 1.6.4 describes how a scenario is briefed, credited and reset. Table 11.3 collects the measured outcomes, and Figure 11.3 draws the rows of cases 3, 6 and 7 as bars.

Table 11.3 Measured outcomes at the default seed

Case	Configuration	Metric	Value
2	Eight at 3.60 psi	mean thickness	700.00 nm
2	—	center to edge	6.94 nm
2	—	*WIWNU*	0.50 %
3	inline map	*RMSE*, wafers 1–3	14.6, 4.4, 0.35 nm
3	lot mean	*RMSE*, wafer 1	16.15 nm
3	product nominal	*RMSE*, wafer 1	29.76 nm
3	feedback only	mean error	+25.0 nm, all
4	fixed recipe, 260 wafers	leaves band at wafer	70
4	—	end-of-life error	+90 nm
5	$\lambda = 0.08$	bias, spread	−15.0, 6.4 nm
5	$\lambda = 0.30$	bias, spread	−10.3, 6.4 nm
5	$\lambda = 0.90$	bias, spread	−9.0, 9.2 nm
5	$\lambda = 0.90$	pressure travel	2.64 psi
6	hybrid APC	wafers in spec	84.2 %
6	feedforward only	wafers in spec	28.2 %
6	feedback only	wafers in spec	24.4 %
6	fixed recipe	wafers in spec	1.3 %
6	$\gamma = 0, 0.15, 0.4$	wafers in spec	80.5, 84.2, 87.1 %
6	$\lambda = 0.9$, $\gamma = 0.15, 0.6$	wafers in spec	89.8, 60.7 %
7	soft reset	qual wafers, worst	1, 2.5 nm
7	carry-all reset	qual wafers, worst	1, 41.7 nm
7	cold reset	qual wafers, worst	6, 59.4 nm
8	adaptive policy	wafers, release	1, 11.5 min
8	fixed policy	wafers, release	6, 18.4 min
8	after one wafer	$\sigma_{\hat{\theta}_g}$	0.0500 to 0.0070
9	VM only, 18 nm bias	surrogate, true	699.6, 682.8 nm
9	—	wafers in spec	46 % to 24 %
10	26 nm azimuthal	radial *RMSE*	0.14–4.8 nm
10	—	total *RMSE*	18.4–19.0 nm
11	as given	wafers in spec	13 of 890
11	suggested	wafers in spec	862 of 914
11	as given, suggested	qual wafers	12, 2
11	as given, suggested	film violations	333, 0
11	as given, suggested	worst mean error	140.4, 7.2 nm

11.6.1 The Plant Is Tall

A profile error is not a list of eight errors. Projecting a radial ripple of four full periods across the radius, $\sin(8\pi\rho)$, onto the column space of G by unregularized least squares leaves 66 % of it unreachable, and a ripple of two periods, $\sin(4\pi\rho)$, leaves 10 %. The right question about a residual is therefore not which chamber to move but whether the shape is in the reachable subspace at all, and the application reports the radial error and the azimuthal floor separately.

> **+ DID YOU KNOW? Equal Pressure, Unequal Removal**
>
> From the simulator: in Table 11.3, case 2 polishes with all eight chambers at 3.60 psi and hits the mean of 700.00 nm, yet the center finishes 6.94 nm thicker than the edge. The pad sweeps the wafer's edge faster than its center (Section 11.3), so equal pressure is not equal removal.

The unreachable case with the cleanest answer is azimuthal. Scenario 10 adds 26 nm of asymmetry with zero radial mean, a floor of $26/\sqrt{2} = 18.38$ nm. Over the 240 min run the radial residual stays between 0.14 and 4.8 nm while the total profile error is 18.4 to 19.0 nm against a 12 nm limit. Raising

λ to 0.8 does not move the floor. It lets the radial part reach 9.0 nm and the total 20.5 nm, because a loop tuned harder against an azimuthal error only adds radial noise to it. Control authority has a shape, and head gimbal, wafer bow, retaining ring wear and pad groove orientation are fixed in the machine shop rather than in the control system.

11.6.2 Feedforward and Feedback Answer Different Questions

Scenario 3 changes product at 90 min, flipping the incoming center-to-edge shape from 0.6 to −0.7 units of tilt (from 24 nm edge-thick to 28 nm center-thick at 40 nm per unit) on a film that stays 25 nm thick. On the first wafer after the change the inline map leaves a profile error of 14.6 nm, the second 4.4 nm and the third 0.35 nm, and after the second change at 180 min 18.9, 7.3 and 1.4 nm. The run-to-run move limit of 0.15 psi per wafer, which exists to guard against a noisy map, also meters out the feedforward correction, and raising it to 0.50 psi brings the first wafer to 1.08 nm. The lot-representative mean leaves 16.15 nm on every wafer and the product nominal 29.76 nm. The lot mean corrects mode zero and nothing else, so what survives it is the shape. Run-to-run feedback alone never removes the 25 nm level error at all, because the residual it learns from is the measured removal and an incoming film error is not a removal error. That is not a tuning failure. The information that would have fixed those wafers existed before they were polished and the architecture had no path for it.

Pause. Try Exercise 11.1 before reading on.

> **TRY IT IN THE APP 3 · Feedforward Sees the Wafer in Front of the Tool**
>
> Load the scenario and press Run. With Incoming profile known at Inline map, as shipped, watch the profile error on the first three wafers after the product change at 90 min: 14.6, 4.4 and 0.35 nm. Press Reset, set it to Lot mean before Run, and run again: 16.15 nm of shape stays on every wafer.

Feedback earns its place against everything upstream metrology cannot see. In scenario 6 a hybrid architecture holds 84.2 % of wafers in specification (255 of 303) across a whole pad life containing a slurry step and a product change. Feedforward alone holds 28.2 % and feedback alone holds 24.4 %, and the two middle numbers are close for entirely different reasons. Feedforward misses the slurry step completely, so every wafer after it is out of specification, and feedback never corrects the incoming shape change at 260 min, after which all 127 of its wafers are out. A fixed recipe runs the flat qualified recipe and holds 1.3 %.

11.6.3 What the Weight Trades, and What the Trend Term Costs

Scenario 5 drifts slurry activity upward in four steps while a poorly matched metrology tool adds 7 nm of correlated error to every map. Over the 229 production wafers polished from the first step at 60 min to 400 min, at $\lambda = 0.08$ the loop carries a standing bias of −15.0 nm with a spread (standard deviation) of 6.4 nm and moves the mean chamber pressure a total of 0.49 psi. At $\lambda = 0.30$ the bias is −10.3 nm at a spread of 6.4 nm and 0.96 psi. At $\lambda = 0.90$ the bias reaches −9.0 nm and costs 9.2 nm of spread, 2.64 psi of actuator travel and a worse peak error, 34.6 nm against 28.5 nm. The bias is the lag against the ramp and the spread is map noise being copied into silicon, the trade of Section 1.4.3. The bias shrinks monotonically across the whole slider, from −20.8 nm at $\lambda = 0.02$ to −9.0 nm at 1.0. The spread is smallest, 6.1 nm, near $\lambda = 0.12$ and grows on both sides, below it because a loop that far behind the ramp lets the four steps through as variance.

The trend weight is worth a few points here, and it turns on its owner only when the level weight is fast as well. Across a pad life at the shipped $\lambda = 0.30$, $\gamma = 0.15$ holds 84.2 % of wafers in specification against 80.5 % at $\gamma = 0$, twelve more good wafers, and the worst mean error in the second half of the pad falls from 12.3 nm to 6.9 nm. Larger trend weights keep paying at this level weight, 87.1 % at $\gamma = 0.4$ and 89.4 % at 0.6, because the two maps of each lot are twelve wafers apart and every slope is read across about twelve wafer-equivalents of exposure. At $\lambda = 0.9$ the level jumps most

of the way to the slurry step on the first map that sees it, the trend term reads that jump as the start of a ramp and extrapolates it across about 23 wafer-equivalents of metrology delay, and the mean overshoots the step. With $\gamma = 0.6$ the overshoot reaches 35.6 nm and the yield falls to 60.7 % (185 of 305), against 89.8 % (273 of 304) at $\gamma = 0.15$. A level step and the start of a ramp look the same across a metrology delay, and the challenge's suggested settings keep γ at 0.05.

11.6.4 The Price of Three Reset Strategies

Scenario 7 replaces a pad on identical hidden parameters and runs each strategy from the same seed. A soft reset releases production after one mapped monitor wafer and the worst of the first twelve production wafers is 2.5 nm off target. A carry-all reset releases just as fast, on a confident belief about a pad that is no longer on the machine, and the worst of the first twelve is 41.7 nm. A cold reset spends six qualification wafers and is still 59.4 nm out, because counting wafers is not the same as learning from them. The soft case at the end of that first production lot is shown in Figure 11.4.

> **TRY IT IN THE APP 7 · The Pad Change, and What the Estimator Keeps**
>
> Load the scenario and press Run. Watch the Qual wafers tile after the pad is replaced at 30 min. Under the shipped Soft reset one mapped wafer releases production, and the worst of the first twelve production wafers is 2.5 nm off target. Press Reset, choose Carry all for Pad-change reset before Run, and run again: the release is as fast and the worst wafer is 41.7 nm off.

The single mapped wafer of the soft reset takes the standard deviation of the global gain estimate from 0.0500 to 0.0070 and the predicted joint pass probability to 0.99989. On the fresh pad of scenario 8 the adaptive policy therefore releases on the wafer polished at 11.5 min, where the fixed count releases at 18.4 min, one wafer against six.

Two qualifications are owed to that result. The correlated whole-map term caps how much any one wafer can say about global gain. Without it the same wafer would take the gain standard deviation 6.5 times lower, while the tilt, curvature and edge spreads (0.0050, 0.0041 and 0.0110 after the wafer) would hardly change. And the D-optimal choice of excitation, which is the right method and costs nothing, is worth well under one percent of the posterior spread on this plant (0.6 % in the geometric mean of the four standard deviations, at most 2.4 % on any one). A multiplicative pad correction is seen almost as well by any recipe that removes material, so the evidence is in the retained prior and in the thirty-three sites rather than in the experiment design. That is a measured result about this class of correction and not a general statement about experimental design.

The gate is also a gate. Triggering a chamber tracking fault before release causes the supervisor to refuse release whatever the probability says, because a chamber that is not tracking makes the whole prediction meaningless. In scenario 8 the tool spends its cap of four qualification wafers and goes to hold at 15.7 min.

11.6.5 A Surrogate With No Anchor

Scenario 9 removes physical metrology entirely and gives the virtual model an 18 nm bias at 40 min. By 100 minutes the surrogate reads 699.6 nm and the wafers are 682.8 nm. The bias enters the run-to-run residual as if it were a process error, the recipe removes it, and the true thickness moves by the whole of it while the measured mean on the top trend stays on target. In a fab the yield record would be the only symptom, and here the share of wafers in specification falls from 46 % at 100 min to 24 % at 190 min. The simulator also draws the actual wafer thickness as a dashed trace, which no real tool has. Physical metrology returns at 200 min with a one-lot delay, so the first physical map arrives at 239.4 min and finds the removal estimate carrying the whole bias. Its gain-mode residual differs from the estimator's prediction for that wafer by 21.5 nm, more than four times the 3.8 nm noise of a physical map, and the application raises its anchor alarm. That map moves the level estimate from −18.4 nm to −12.0 nm, and the actual mean is back above 692 nm at 241 min.

The return is that fast, and then too fast, because of the trend term of Section 11.4.4 at its default $\gamma = 0.12$. The level recovering from −18.4 nm looks exactly like the start of a ramp, the slope is 0.37 nm per wafer-equivalent after the first map and still 0.19 nm per wafer-equivalent near 387 min, and the estimator extrapolates it across the metrology delay. The actual mean overshoots to 711.0 nm at 350 min and is 2.4 to 10.4 nm high between 350 and 400 min. With $\gamma = 0$ the peak is 702.4 nm and the mean from 350 to 400 min is 699.0 nm, at the price of a slower return, above 692 nm only at 270.5 min.

The surrogate in the application is deliberately anchored at the last physical map rather than carrying the controller's live residual estimate. Letting it carry the live estimate makes the loop feed on its own output, the level then grows by λ times the bias at every update, the minimum film limit is crossed eleven wafers after the bias appears, and the mean settles near 450 nm once the level reaches its bound. That is a different and much less instructive failure than the one the scenario is about. The soft sensor of Volume 5, Chapter 10 is kept honest the same way, by a sampled and delayed laboratory that corrects its bias.

11.6.6 The Graded Lifecycle

The challenge runs 1400 minutes containing a slurry step, two product changes, one corrupted metrology map and two pad replacements, starting from a fixed recipe, a carry-all reset, a fixed six-wafer qualification policy and an EWMA weight of 0.06. It is scored on actual production thickness, not on the measured trend. The worst deviation of the production-wafer mean from 700 nm from 924 min to the end must stay within ±10 nm, and no more than six qualification wafers may be spent across both replacements. At least one full lot of 25 production wafers must be completed in the graded window, and the tool must finish released for production. This prevents a held last-wafer value from satisfying the thickness target without making product. The challenge is judged on the case as given, so changing a model parameter or a disturbance, including firing an extra disturbance button, voids the attempt. The suggested run completes 323 production wafers in that window. As given, the worst deviation is 140.4 nm, 13 wafers of 890 meet specification, twelve qualification wafers are consumed and the minimum film limit is crossed 333 times. With the suggested settings, which correct the architecture, the reset strategy, the qualification policy and the estimator weights, the same seeded run gives a worst deviation of 7.2 nm, 2.8 nm inside the bound, 862 of 914 wafers in specification, two qualification wafers in total, no film violations and 39 min of downtime. In the same graded window the measured mean on the top trend strays as far as 7.3 nm from target, which is metrology noise and delay rather than product. The corrupted map is delivered at 754.1 min with a normalized innovation of 9.34 and is rejected, just above the threshold of nine.

Most of that difference is the architecture and the estimator weights. The suggested settings with the shipped carry-all reset and fixed six-wafer qualification put back still hold 829 of 906 wafers in specification, but they spend twelve qualification wafers and miss the mean by 19.7 nm. The shipped architecture with only the reset and the qualification policy corrected holds 13 of 900. The two replacements decide the qualification budget, and the architecture decides the yield. No single change from the shipped setup passes. A pass in production takes the inline map read by Feedforward or Hybrid APC, a soft reset and at most three qualification wafers per pad, and with those alone the shipped estimator weights and metrology settings pass, at 4.8 nm and 85 % in specification.

11.7 Technology in Practice

The layers of Section 11.4 run on a thickness gauge, eight pressure regulators, the polish timer and the conditioner, and a recipe controller above the tool. Figure 11.5 draws them around the polisher of Figure 11.1.

What Measures It

The model reduces the gauge to a map of thirty-three radii on two wafers a lot, with 2.0 nm of independent noise at each site and 3.0 nm of correlated error on the whole map, delivered one or two

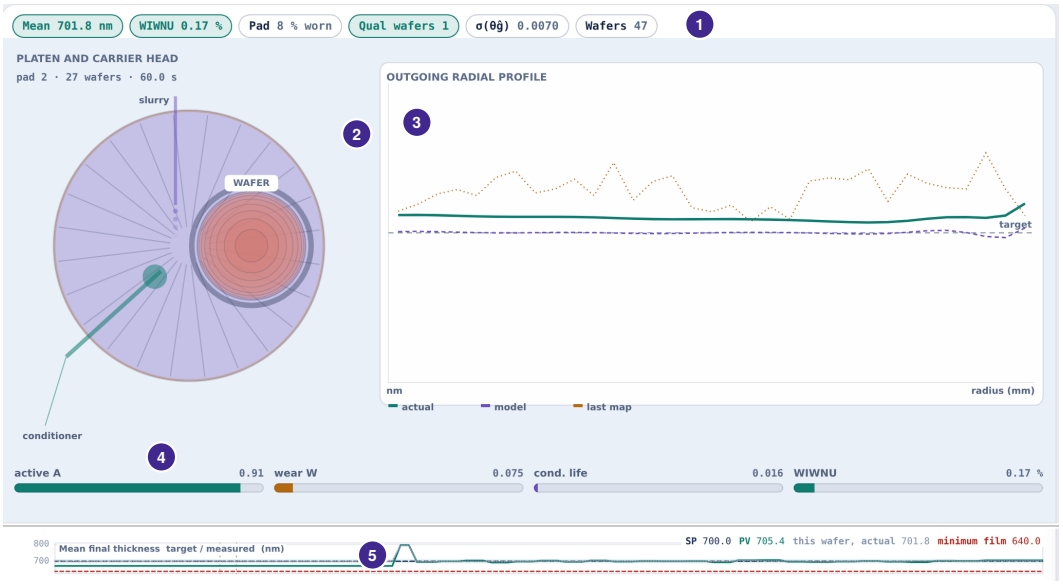


Figure 11.4 The application during scenario 7 at 80 min, 50 min after a pad replacement under the soft reset strategy and at the end of the first production lot on the new pad, cropped to the plant view and the top trend. (1) The badges read one qualification wafer and a standard deviation of the gain estimate of 0.0070. (2) The scene shows the platen from above with the slurry feed and the conditioner arm, and the carrier head chambers colored by the pressure they are holding. (3) The outgoing radial profile of the last wafer is drawn against the model prediction and the last delivered map. (4) The bars give the pad active surface (0.91), structural wear (0.075) and conditioner life. (5) The trend over the last hour carries the measured mean thickness with the actual wafer beside it (705.4 and 701.8 nm at this instant) and the minimum film limit. The single qualification wafer polished at 40.5 min stands out as the deliberately under-polished excursion to 792 nm.

lots late. Inside a tool, endpoint systems watch the polish while it happens. The motor current method is the less expensive, but its slow response limits what it can show of the process [17]. The chapter's polisher has no such sensor, so it learns the removal only from a map.

> **TECHNOLOGY** **Sensor: An Optical Map of Film Thickness**
>
> Optical methods that measure a film's thickness, the light it scatters or the reflectance of the wafer surface are the most common way to see a polished wafer. They give direct information about the surface, because the signal depends little on the other polishing conditions, while the pad itself is hard to watch this way [17]. The 1996 run-by-run controller read nine points on each blanket wafer [8], where this chapter reads thirty-three radii. No open, vendor-neutral source gives a production gauge's range or accuracy, so none is given here. A gauge that is poorly matched to its fleet adds the same error to every site of a map, as the 7 nm of case 5 does.

The inline pre-polish map is also a choice with a price: feedforward conventionally needs every lot measured, which most operations cannot afford [18].

What Moves It

The model's eight chamber pressures run from 2.5 to 4.8 psi, each a first order lag of 3.0 s at the center chamber falling to 0.5 s at the edge, beside a polish time of 35 to 90 s and a conditioning dose of 0 to 1.5.

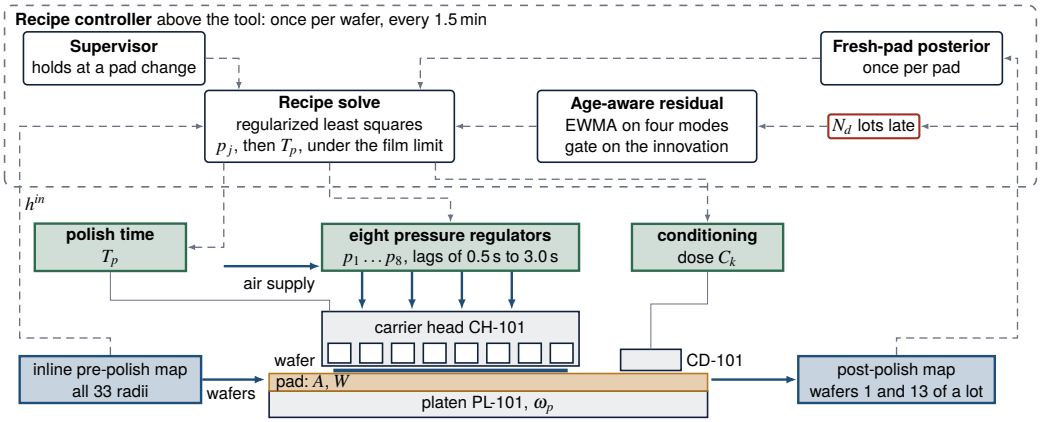


Figure 11.5 The layers of the simulated polisher drawn with their hardware. Eight pressure regulators feed the chambers of carrier head CH-101 and settle in seconds inside one wafer. The recipe solve writes their setpoints, the polish time T_p and the conditioning dose C_k once per wafer, from the inline pre-polish map and the controller's model. The post-polish map returns one to two lots late to the age-aware residual, and the fresh-pad posterior updates once per pad. The supervisor holds production at a pad change. Notice that the polished film is measured only after the next lots are already polished.

TECHNOLOGY **Final Element: An Air Pressure Behind Each Ring**

Each chamber of a membrane carrier head is a volume of air whose pressure presses one ring of the wafer against the pad. An electronic controller sets such a pressure through a current-to-pressure transducer, which takes the 4 to 20 mA output signal and acts as an electrically controlled air pressure regulator, reducing the supply pressure to the commanded level [19]. A study of a three-zone head gave each chamber a dynamic element set by the gas it holds [4], and the first-order lag of this chapter's chambers stands for it. A chamber that leaks shows as a pressure that no longer tracks its setpoint, and in case 8 the supervisor then refuses release.

The model's lag, paid for by Equation (11.5), stands for the regulator and the chamber volume. It has no dead band and no supply limit, and its only failure is the tracking fault of case 8.

Where the Controller Runs

In a fab the recipe layer belongs to advanced process control, which combines model-based run-to-run control with fault detection and classification, and run-to-run control adjusts the recipe lot by lot or wafer by wafer [18]. Supervisory and lower-level loops run with cycle times from milliseconds to minutes [20]. The app integrates every 0.05 min with the controller on the same scan, while the recipe and the estimator act once per wafer, every 1.5 min, the residual when a map arrives, and the posterior once per pad.

IN PRACTICE **Breaking In a New Pad**

A new pad is broken in before it polishes product. In one study about 50 copper wafers were polished to break in a pad, the radius of its surface asperities grew from 16 to 93 µm, and scratching became less severe [21]. In another, the surface texture that diamond disc conditioning cut into new pads settled within 30 minutes in four cases of five and within 60 minutes in all [22]. Maintenance shifts a tool's bias in a step [18], and the industry's usual answer is a fixed count of sacrificial qualification wafers after every pad change (Section 11.1).

In such a system the chapter's structure is a constrained least squares recipe block, an estimator for each of four profile modes behind an innovation gate, a Bayesian update of the fresh-pad parameters, and the supervisor's state machine.

What Is Modernizing It

The first development is virtual metrology with a stated confidence. A regression whose predictive variance sets the weight of the run-to-run EWMA did better on a polishing case study than schemes that ignore the reliability of the prediction [15], and models trained on production data from a high-volume fab predict non-uniformity and the removal across the wafer, with polishing time and carrier rotation among the strongest factors [14]. The second is sensing in the tool: a force transducer, the motor current and an acoustic emission sensor see different bands of the process and complement one another for monitoring and endpoint detection [17]. A recent review names prediction driven by artificial intelligence, sensing in the tool and cleaner consumables as what is turning polishing into a data-informed process [23]. Case 9 of this chapter shows the condition attached to the first.

11.8 Limitations

The influence matrix is a compact empirical representation. It omits membrane and pad mechanics, slurry hydrodynamics, pad groove transport, wafer bending, retaining ring contact mechanics and detailed relative velocity trajectories, and a real identified matrix must be loaded rather than generated.

The application grades a radial blanket-film profile. It does not predict die scale erosion, dishing, defects, scratches, residue, corrosion or electrical yield, and the eight chambers and their boundaries are supplied values for this project rather than an industry standard.

The release probability is only as meaningful as the covariance calibration behind it, and the joint probability multiplies three marginals that do not fail independently. It supplements process qualification and statistical process control and replaces neither.

Virtual metrology here is a model output with an explicit uncertainty, a switchable bias and an anchor at the last physical map. A real deployment needs data governance, product and tool matching, drift detection and independent validation, none of which is modeled.

OPEN CHALLENGE **A Virtual Measurement Worth Trusting**

A fab generates millions of data points a year for process control, and models trained on polisher data already predict non-uniformity and the removal at each point of a wafer [14]. A 2025 review of machine learning in polishing names what is still hard: data leakage, sparse target labels and hybrid modeling. It calls for benchmark data sets and for models of dishing and within-wafer non-uniformity [24]. Case 9 shows the stake. A surrogate with no physical anchor carries its own bias into the product. Progress would be a prediction whose confidence stays honest across a pad change, so that it can set how often wafers are measured [16].

The D-optimal choice of qualification recipe is measured at under one percent of the posterior spread on this plant. That is a property of a multiplicative pad correction observed through a shared aperture and should not be carried to a plant where the parameters enter differently.

The economic values are teaching assumptions. Wafer value, scrap cost, qualification cost, pad cost and downtime cost are round numbers chosen so the trade-offs are visible and are not fab cost data.

11.9 Conclusion

This chapter studied a 300 mm polisher as an interactive simulator in which eight carrier head chambers drive thirty-three radial outputs while the pad doing the work wears out and is eventually replaced. Eight continuous states carry the consumable and the unmeasured disturbances, a constrained least squares recipe solve carries the spatial inverse with its smoothness and gradient limits, an age-aware residual estimator carries the drift, and a four parameter Bayesian update carries the pad-to-pad variation.

The measured outcome that organizes the chapter is the price of a maintenance event. On identical hidden pads, a soft reset releases production after one mapped monitor wafer and lands 2.5 nm off target, a carry-all reset releases just as quickly on a confident belief about a pad that is in the bin and lands 41.7 nm off, and a cold reset spends six qualification wafers and is still 59.4 nm out. The rule that separates them is short enough to remember. Reset what belongs to the part that was replaced, keep what belongs to the machine, and raise the uncertainty rather than the gain.

KEY POINTS

- A polishing pad wears with every wafer and is replaced every few days, and the controller must then decide what it still believes.
- Eight concentric chambers cannot reach every radial profile, and no concentric chamber can touch an azimuthal error.
- Feedforward and feedback answer different questions: hybrid control held 84.2 % of wafers in specification, feedforward alone 28.2 % and feedback alone 24.4 %.
- Reset what belongs to the part that was replaced, keep what belongs to the machine, and raise the uncertainty rather than the gain.

Where else this applies. A regularized inverse of an ill-conditioned response in Volume 6, Chapter 6, a capacity that cleaning restores only in part in Volume 8, Chapter 11, and a soft sensor kept honest by a delayed laboratory in Volume 5, Chapter 10.

References

[1] K. S. Gokhale and B. M. Moudgil, “Particle technology in chemical mechanical planarization,” *KONA Powder and Particle Journal*, vol. 25, pp. 88–96, 2007.

[2] E. Sachs, A. Hu, and A. Ingolfsson, “Run by run process control: combining SPC and feedback control,” *IEEE Transactions on Semiconductor Manufacturing*, vol. 8, no. 1, pp. 26–43, 1995.

[3] E. Del Castillo and A. M. Hurwitz, “Run-to-run process control: Literature review and extensions,” *Journal of Quality Technology*, vol. 29, no. 2, pp. 184–196, 1997.

[4] S.-J. Shiu, C.-C. Yu, and S.-H. Shen, “Multivariable control of multizone chemical mechanical polishing,” *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 22, no. 4, pp. 1679–1687, 2004.

[5] J. Luo and D. A. Dornfeld, “Material removal mechanism in chemical mechanical polishing: theory and modeling,” *IEEE Transactions on Semiconductor Manufacturing*, vol. 14, no. 2, pp. 112–133, 2001.

[6] W.-T. Tseng and Y.-L. Wang, “Re-examination of pressure and speed dependences of removal rate during chemical-mechanical polishing processes,” *Journal of The Electrochemical Society*, vol. 144, no. 2, pp. L15–L17, 1997.

[7] E. Paul, “A model of chemical mechanical polishing,” *Journal of The Electrochemical Society*, vol. 148, no. 6, p. G355, 2001.

[8] D. S. Boning, W. P. Moyne, T. H. Smith, J. Moyne, R. Telfeyan, A. Hurwitz, S. Shellman, and J. Taylor, “Run by run control of chemical-mechanical polishing,” *IEEE Transactions on Components, Packaging, and Manufacturing Technology: Part C*, vol. 19, no. 4, pp. 307–314, 1996.

[9] A. Chen and R.-S. Guo, "Age-based double EWMA controller and its application to CMP processes," *IEEE Transactions on Semiconductor Manufacturing*, vol. 14, no. 1, pp. 11–19, 2001.

[10] J. Moyne, E. Del Castillo, and A. M. Hurwitz, Eds., *Run-to-Run Control in Semiconductor Manufacturing*. Boca Raton, FL, USA: CRC Press, 2001.

[11] Y. H. Sun, R. K. Kang, and D. M. Guo, "Modeling of back pressure distribution on the wafer loaded in a multi-zone carrier in chemical mechanical polishing," *Materials Science Forum*, vol. 532-533, pp. 233–236, 2006.

[12] M. Tsujimura, "Chemical mechanical polishing (CMP) removal rate uniformity and role of carrier parameters," in *Advances in Chemical Mechanical Planarization (CMP)*. Elsevier, 2016, pp. 417–432.

[13] Z. C. Li, E. A. Baisie, X. H. Zhang, and Q. Zhang, "Diamond disc pad conditioning in chemical mechanical polishing," in *Advances in Chemical Mechanical Planarization (CMP)*. Elsevier, 2016, pp. 327–357.

[14] M. Breidung, T. Rothe, A. Lauff, P. Thieme, J. Langer, M. Günther, and H. Kuhn, "Process data-driven machine learning for non-uniformity prediction and virtual metrology in chemical mechanical planarization," *Journal of Intelligent Manufacturing*, 2025.

[15] J. Wan and S. McLoone, "Gaussian process regression for virtual metrology-enabled run-to-run control in semiconductor manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 31, no. 1, pp. 12–21, 2018.

[16] G. A. Susto, "A dynamic sampling strategy based on confidence level of virtual metrology predictions," in *2017 28th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*. IEEE, 2017, pp. 78–83.

[17] H. Jeong, H. Kim, S. Lee, and D. Dornfeld, "Multi-sensor monitoring system in chemical mechanical planarization (CMP) for correlations with process issues," *CIRP Annals*, vol. 55, no. 1, pp. 325–328, 2006.

[18] A. V. Prabhu, "Performance monitoring of run-to-run control systems used in semiconductor manufacturing," Ph.D. dissertation, The University of Texas at Austin, 2008. [Online]. Available: http://hdl.handle.net/2152/17721

[19] T. R. Kuphaldt, *Lessons in Industrial Instrumentation*, version 2.33 ed. Open textbook, Creative Commons Attribution 4.0 license, ibiblio.org/kuphaldt/socratic/sinst, 2024.

[20] K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, A. Hahn, S. Saravia, A. Sherule, and M. Thompson, "Guide to operational technology (OT) security," National Institute of Standards and Technology, Gaithersburg, MD, Special Publication 800-82r3, 2023.

[21] T. Eusner, N. Saka, and J.-H. Chun, "Breaking-in a pad for scratch-free, Cu chemical-mechanical polishing," *Journal of The Electrochemical Society*, vol. 158, no. 4, p. H379, 2011.

[22] J. McAllister, C. Stuffle, Y. Sampurno, D. Hetherington, J. Sierra Suarez, L. Borucki, and A. Philipossian, "Effect of conditioner type and downforce, and pad break-in time, on pad surface micro-texture in chemical mechanical planarization," *ECS Journal of Solid State Science and Technology*, vol. 7, no. 11, pp. P677–P688, 2018.

[23] H. Je, S. Kang, and S. Kim, "Challenges and innovations in chemical mechanical polishing in the more-than-moore era," *International Journal of Precision Engineering and Manufacturing-Green Technology*, vol. 13, no. 5, pp. 2717–2756, 2026.

[24] G. Winkler, T. Rothe, M. A. Sayyed, L. Jäckel, J. Langer, H. Kuhn, and M. Stoll, "Machine learning in chemical-mechanical planarization: A comprehensive review of trends, applications, and challenges," *Advanced Engineering Informatics*, vol. 68, p. 103663, 2025.

11.10 Exercises

11.1 ⌕ Concept check · 2 min

In scenario 3 the incoming film is 25 nm thick. With Recipe authority at EWMA R2R the 25 nm level error stands on every wafer of the run, while Hybrid APC with the inline map removes it. Why does run-to-run feedback alone never correct it?

(a) Its residual is measured removal minus predicted removal, and an incoming film error is not a removal error.

(b) The metrology delay holds every map back until after the run ends, so no correction can arrive in time.

(c) The run-to-run move limit of 0.15 psi per wafer meters each correction down to nothing.

(d) An EWMA weight of 0.30 is too small to follow a step of 25 nm within the run.

11.2 Concept check · 2 min

In scenario 7 the carry-all reset releases production after one qualification wafer, as fast as the soft reset, yet the worst of the first twelve production wafers is 41.7 nm off target against 2.5 nm for the soft reset. Why does the release gate pass under carry-all?

(a) Carry-all switches the qualification policy to a fixed count of one wafer.

(b) The gate is computed from the controller's own belief, which keeps the worn pad's estimate with a narrowed covariance.

(c) The qualification wafer under carry-all carries no excitation, so its map looks like a normal wafer.

(d) The first production wafers are measured two lots late, so the gate cannot see their error.

11.3 Concept check · 1 min

A step of 0.20 psi on interior chamber 4 stays above half its peak across 21 % of the wafer radius, while the same step on edge chamber 8 peaks lower, at 0.081 psi, and stays above half of it across 3 %. What gives the edge chamber its localized authority?

(a) It has a wider pressure range than the interior chambers.

(b) It has the fastest pressure time constant of the eight.

(c) Its radial band is narrow, so its footprint in G is narrow.

(d) It acts together with the retaining ring, which doubles its effect.

11.4 Theory and calculation · 6 min

The platen turns at 80 rpm, the carrier at 60 rpm, the centers are $d = 0.220\,\mathrm{m}$ apart and $R = 150\,\mathrm{mm}$. The innermost calculation point is at $\rho = 0.5/33$ and the outermost at $\rho = 32.5/33$. Use Equation (11.3) and give speeds to four decimals.

(a) Find the relative speed at the innermost and outermost points and their ratio. *Checkpoint:* 1.8431 m/s, 1.8689 m/s, ratio 1.0140

(b) With $\beta = 1$, eight equal pressures, a uniform pad and a mean removal of 500 nm, removal at each point is $500\,v_m/v_{ref}$ nm. Take $v_{ref} = 1.8564\,\mathrm{m/s}$ and find how much thicker the center finishes than the edge, to 0.01 nm. Compare with scenario 2. *Checkpoint:* 6.94 nm

(c) Which carrier speed makes the relative speed the same at every radius? What does scenario 2 then show for the center and edge thickness? *Checkpoint:* 80 rpm; both 700.00 nm

11.5 Theory and calculation · 8 min

Pad wear lowers removal by the factor $1 - d_w W^{1.4}$ with $d_w = 0.18$ (Equation (11.10)). Scenario 4 runs a fixed recipe that removes 500 nm on a pad of $N_{pad} = 260$ wafers at the default dose $C_{nom} = 0.55$, so wafer k is polished at $W = (k-1)/260$. Ignore the slow loss of conditioner efficiency.

(a) Find the removal factor at end of life, $W = 1$, and how far above target the mean then finishes. *Checkpoint:* 0.82, 90 nm

(b) Find the wear at which the mean is 14 nm above target, to four decimals, and the first wafer outside the ± 14 nm band. *Checkpoint:* $W = 0.2647$, wafer 70

(c) The dose is raised to $C = 1.00$ on a new conditioner ($\eta_{cond} = 1$). With one wafer-equivalent of exposure per wafer, find the settled active surface A^* at both doses (four decimals) and the pad life in wafers at $C = 1.00$ from Equation (11.8), to one decimal. *Checkpoint:* 0.9083 and 0.9474; 236.6 wafers

11.6 Theory and calculation · 7 min

A qualification wafer is mapped at $M = 33$ sites with $\mathbf{R} = s^2\mathbf{I} + \sigma_c^2\mathbf{1}\mathbf{1}^T$, $s^2 = \sigma_{ind}^2 + \sigma_{proc}^2$, $\sigma_{ind} = 2.0\,\text{nm}$, $\sigma_{proc} = 1.2\,\text{nm}$ and $\sigma_c = 3.0\,\text{nm}$. Treat the gain θ_g alone, with a prior standard deviation of 0.05 and a sensitivity $\mathbf{H} = r\mathbf{1}$ with $r = 430\,\text{nm}$ at every site (86 % of the nominal 500 nm removal).

(a) Use Equation (11.17) to show that $\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} = Mr^2/(s^2 + M\sigma_c^2)$. Evaluate it and the posterior standard deviation of θ_g to four decimals. *Checkpoint:* about 20 175; 0.0070

(b) Repeat with $\sigma_c = 0$, and give the ratio of the two standard deviations to one decimal. *Checkpoint:* 0.00094; 7.4

(c) How many independent sites, each of variance s^2, would carry the same information about θ_g as the 33 correlated ones (two decimals)? What standard deviation does a wafer with infinitely many sites reach (four decimals)? *Checkpoint:* 0.59 sites; 0.0069

11.7 App challenge · 8 min

Load scenario 7, "The pad change, and what the estimator keeps", at the default seed. Run each case to 80 min at the default speed of 20 min per second and export the CSV, whose time column is `Time (h)`.

(a) As shipped (Pad-change reset at Soft), read the Qual wafers tile, the `est_sigma_gain` column at 0.70 h, and the largest $|\texttt{mean_prod_nm} - 700|$ in the rows from 0.68 to 0.99 h, to 0.1 nm. *Checkpoint:* 1 wafer; 0.0070; 2.5 nm

(b) Press Reset, choose Carry all before Run, and read the same three values. *Checkpoint:* 1 wafer; 0.0100; 41.7 nm

(c) Press Reset, choose Cold before Run, and read them again. Explain the three values of `est_sigma_gain`. *Checkpoint:* 6 wafers; 0.0500; 59.4 nm

11.8 App challenge · 8 min

Load scenario 8, "The qualification wafer sprint", at the default seed. Run each case to 30 min at the default speed and export the CSV (time column `Time (min)`). In each part find the first row whose `supervisor_state` is `enhanced` or `hold`, and read its time, `qual_total`, `est_sigma_gain` (four decimals) and `pass_prob` (five decimals).

(a) As shipped, with Qualification policy at Adaptive. *Checkpoint:* 12.0 min, enhanced; 1; 0.0070; 0.99992

(b) Press Reset and choose Fixed count before Run. *Checkpoint:* 18.5 min, enhanced; 6; 0.0029

(c) Press Reset, set the policy back to Adaptive and Release probability (in Advanced) to 0.9999, its editable maximum, before Run. *Checkpoint:* 12.0 min, enhanced; 1

(d) Reload the scenario, press Run, and press Zone 6 develops a leak at any time before 11.5 min. *Checkpoint:* 16.0 min, hold; 4

11.9 App challenge · 8 min

Load scenario 10, "The error concentric zones cannot reach", at the default seed, run it to 240 min at the default speed of 20 min per second, and export the CSV. Give values to 0.01 nm.

(a) Read the smallest and largest `rmse_nm` (the radial part), the largest `rmse_total_nm`, and `uncorrectable_nm`. Check the largest total against the sum in quadrature. *Checkpoint:* 0.14 and 4.81 nm; 19.00 nm; 18.38 nm

(b) Reload, set EWMA level weight to 0.80 before Run, and read the largest radial and total values again. *Checkpoint:* 9.03 nm; 20.48 nm

(c) The floor is $a/\sqrt{2}$ for an Azimuthal asymmetry of a nm peak, set in steps of 1 nm. Using the largest radial value of part (a), predict the largest a for which every wafer meets the 12 nm limit. Reload, set that value and one more before Run, and read the Wafers in spec tile at 240 min. *Checkpoint:* 15 nm: 100 %; 16: 99 %

11.10 Application area · 4 min

A fab runs pads of 1500 wafers. Its qualification policy is a fixed count of six monitor wafers after every pad change.

(a) What percentage of all wafers polished go to qualification under the fixed count, and under an adaptive policy that releases after one mapped wafer, as in scenario 8? Give two significant figures. *Checkpoint:* 0.40 % and 0.067 %

(b) The fab changes pad supplier. What should happen to the fresh-pad prior and to the release decision, and what does the simulator not do for the engineer? *Checkpoint:* widen or set aside the prior; nothing detects the change

11.11 Application area · 3 min

In scenario 9 virtual metrology is the only measurement and develops an 18 nm bias at 40 min. By 100 min the surrogate reads 699.6 nm while the wafers are 682.8 nm.

(a) Why does the measured mean on the top trend stay on target? *Checkpoint:* the loop removes the bias as if it were a process error

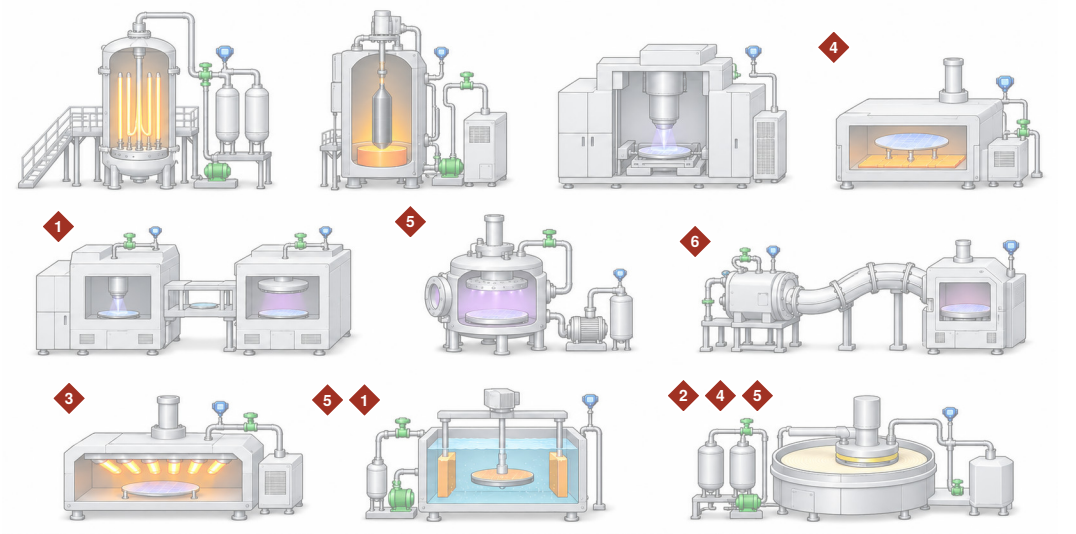
(b) The first physical map disagrees with the estimator by 21.5 nm. The physical map noise is 3.8 nm and the alarm threshold is four times that. Find the ratio to one decimal, and say how a production deployment should use such an alarm. *Checkpoint:* 5.7; raise the physical sampling rate

11.12 Review the agent · 5 min

An AI assistant reviews scenario 6, “Hybrid APC across a pad life”, and writes: “At the shipped level weight $\lambda = 0.30$, raising the trend weight γ from 0.15 to 0.6 lifts the wafers in specification from 84 to 89 %. Raising λ to 0.90 at $\gamma = 0.15$ lifts it to 90 %. Both changes help, so set $\lambda = 0.90$ and $\gamma = 0.60$ together for the best tool.” The two measurements are correct. Find the flaw and say how to check it on the app.

CHAPTER 12

Summary and Outlook



Open gaps 1 queues that move 2 metrology unsure 3 unknown film stack 4 sparse maps 5 maintenance events 6 reading validity

Notice that the polisher carries three gaps, and that no gap marks the silicon reactor, the puller or the scanner. Illustration, not to scale.

IN PLAIN TERMS **What Ten Wafer Tools Taught**

The main idea. Each chapter of this book ran one tool of a chip factory in a simulator. On almost none of them could the controller see the wafer while the work was being done. A sensor read the heater or the gas, and the wafer was measured later, often after more wafers had gone through.

The challenge. A loop can hold its reading perfectly while the wafer drifts away. A late reading also limits how hard a loop may correct.

Why it matters. The tools that did best saw the cause before it reached the wafer, ended each step on a sign from the wafer itself, and checked their readings against the finished product.

Ten chapters controlled ten tools of a wafer fab, and on almost none of them could a controller see the product while it was being made. A pyrometer read the rod surface while the limit sat at its center, a level sensor read the wafer ahead of the slit, and a microscope read the etched line two lots late. Table 1.1 of Chapter 1 listed, for each tool, what is measured in the tool, what is measured on the product and when, and how the next wafer or lot is corrected. This chapter adds what each chapter measured when the correction worked and when it failed, draws a short guide for choosing a method, looks at where practice is heading, reports how the volume itself was made with AI agents and what they got wrong, and lists the problems that remain open.

12.1 What the Chapters Showed

Tables 12.1 and 12.2 compare the chapters. The second column condenses Table 1.1: what the tool measures and when the product is measured. The other columns name a correction that worked, one

Table 12.1 What Each Tool Measured, and How the Next Wafer Was Corrected, Chapters 2 to 6

Chapter	Measured, and when	What worked	What failed	Measured result
2, polysilicon rods	Surface pyrometer, supply volts and amperes. The center limit is never measured.	A gain schedule on the diameter with an override on the center inferred from the supply.	Fixed gains on a gain that falls 126 times. The schedule alone, which melts the rod.	IAE 760.4 fixed, 43.1 scheduled and 38.8 °C h with both. Melted at 149.3 mm without the override, center peak 1323 °C with it (Table 2.5).
3, crystal puller	Camera diameter, heater power. No instrument reports the pull-rate limit.	Diameter on the pull rate with derivative action, and the heater returning the pull rate to mid-range.	The heater held fixed, so the pull rate drifts toward its limit.	Worst margin 0.221 against 0.358 mm/min, prime crystal 73.46 against 88.00 kg. On a degraded camera, lost at 483.0 min against complete (Table 3.5).
4, wafer scanner	Wafer surface ahead of the slit, before exposure. Overlay on sampled fields after.	Preview feedforward with a lead equal to scan speed times the stage's lag.	A trailing sensor with feedback at any gain. A faster scan without rematching the lead.	Focus RMS 119.0 trailing, 41.6 at the best trailing setting, 13.7 nm with preview. Fields failed 81.7 against 0 (Table 4.5).
5, bake plate	A hundred zone thermocouples every 0.1 s. Linewidth map of every wafer after.	A conductance model as feedforward, a power pulse at each landing, offsets per zone with a neighbor average.	Gains tuned on one zone. Offsets per zone without the average.	All-zone overshoot 18.8 % (Table 5.5). Landing error 4.015 against 0.110 nm. Bowed wafers 3.765 against 0.595 nm (Table 5.4).
6, litho and etch line	Developed line at once. Etched line two lots late.	Feedforward from the developed line. A matched-lot estimator.	Feedback alone at any weight. The newest-reading estimator at a weight of 0.95.	Spread 1.476 with no control, 1.852 with feedback, 0.850 nm with feedforward. Newest reading 8.002 against 1.548 nm matched (Table 6.4).

that failed, and the measured result with the table of the chapter that reports it.

The columns make five points, and Figure 12.1 draws one pair of results from each row.

What can be seen before it arrives is fed forward. The largest single gains of the volume came from corrections made before the product was affected. The scanner's preview cut the focus error to a third of the best that trailing feedback could reach, 13.7 against 41.6 nm. Feedforward from the developed line took the etched spread from 1.852 to 0.850 nm with the feedback tuning untouched, while the feedback alone gave more spread than no control at all. The copper bath dosed by charge voided no wafer on a doubled load that feedback alone met after 97, and the bake plate's landing pulse cut a 4.015 nm linewidth error to 0.110. Each of these disturbances was measured or known in advance, and the feedback loop behind it was left a trim.

The queue and the estimator set the usable gain. A late reading makes a recipe loop a loop with dead time, and Section 1.4.4 derived the boundary. In Chapter 6 the newest-reading estimator at a weight of 0.95 gave a spread of 8.002 nm, and the same weight with the matched-lot estimator gave 1.548. Through a wet clean at a weight of 0.40, a queue of four lots gave 3.812 against 1.497 at two. In Chapter 10 the 42 min age of the analysis set a trim limit of 210, and gains of 300 and 400 voided 16 and 37 wafers where 30 voided none. The limit belongs to the age of the reading and to how the estimator matches it with its lot, not to the plant.

Table 12.2 What Each Tool Measured, and How the Next Wafer Was Corrected, Chapters 7 to 11

Chapter	Measured, and when	What worked	What failed	Measured result
7, plasma etch	Pressure, density, bias, wafer temperature and emission in the step. Seven criteria on the finished wafer.	Ending the main etch on the emission endpoint.	A timed recipe after a wet clean, whether the loops are retuned, closed or both.	Wafers good: 0 of 8 for every change to the loops, 7 of 7 with the endpoint (Table 7.7).
8, ion source	Extraction current every 1 ms. Dose from the scan ledger.	A gate that marks the reading invalid during a spark and holds the command. A local repaint.	A PI loop that chases the spark. A repaint of the whole pass.	Overshoot 14.75 against 0.29 %. Dose error −0.12, +16.58 and +0.21 % for no, whole-pass and local repaint (Table 8.3).
9, spike anneal	Five pyrometers every 10 ms. Sheet resistance after each wafer.	Inverse-model feedforward, a decoupler, a film-stack emissivity table, soak time corrected wafer to wafer.	Feedback alone on the ramp. Five independent loops. One emissivity for every film stack.	Lag 65.2 against 2.9 K. Radial difference 34.8 against 2.9 K. Sheet resistance 26.31 % high on a capped wafer (Table 9.5). Graded lot 0 against 25 of 25 (Table 9.6).
10, plating bath	Charge and temperature always. Additives every 2 h, 42 min old. Thickness each wafer.	Dosing by charge with a sampled trim, a bleed on the modeled breakdown products, a dump.	Feedback alone on the analysis. A trim gain above the limit of 210.	Doubled load: 97 voided wafers on feedback alone, none with the dose by charge. Trim gain 300 and 400: 16 and 37 voids (Table 10.5).
11, polisher	Chamber pressures, pad counters. Removal map one or two lots late, incoming map.	A constrained recipe, incoming feedforward, an age-aware EWMA, a soft reset at a pad change.	Feedback or feedforward alone. Carrying every estimate across a pad change.	In specification 84.2 hybrid, 28.2 feedforward, 24.4 % feedback. After a pad change 2.5 soft against 41.7 nm carried (Table 11.3).

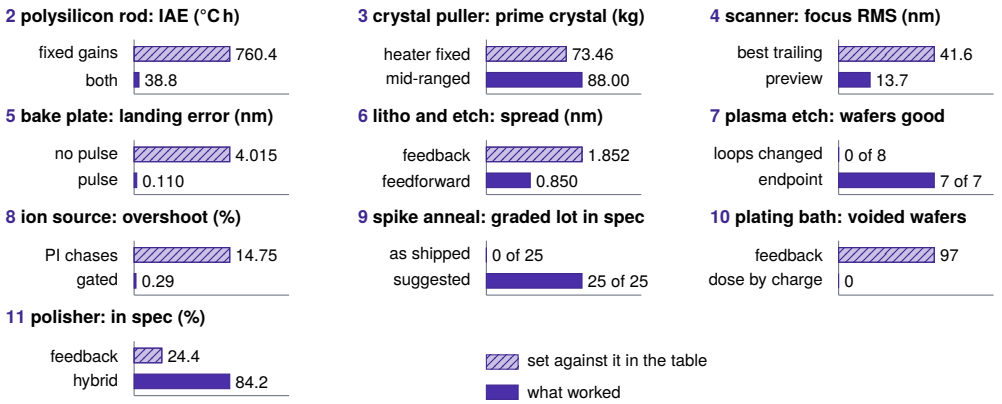


Figure 12.1 One pair of results from each chapter, with the numbers as the "Measured result" column of Tables 12.1 and 12.2 prints them. Each panel has its own scale from zero. A larger value is better in the panels of Chapters 3, 7, 9 and 11. Notice that in every panel the better result came from a change of structure or measurement, not of tuning alone.

A step ends on what the wafer does. After a wet clean the main etch a fresh chamber needs varied by about twelve percent from wafer to wafer. No change to the four chamber loops saved a wafer of the graded cassette, and the emission endpoint saved all seven. The polysilicon batch ends on its center temperature, which no instrument reads. Under the schedule alone the rod melted at 149.3 mm, and the override on the inferred center reached 150 mm with the center at 1323 °C. The ion source repaints the dose a spark interrupted. A local repaint left 0.21 % of dose error where repainting the whole pass left 16.58 %. The step's end, not its loops, decided these results.

Zones on one wafer move together. Section 1.4.5 showed that a loop tuned on one zone sees a smaller gain than the common move. The bake plate's common move had 4.43 times the single-loop gain and overshot 18.8 %. The anneal's five independent loops let the radial difference reach 34.8 K and slip the wafer, and the decoupler held it to 2.9. A hundred run-to-run offsets cut the bowed-wafer signature from 3.765 to 0.745 nm and roughened the map to 0.254, and a neighbor average reached 0.595 and 0.189. Many actuators on one wafer need a model of how they share it.

A reading can be steady and wrong. The anneal's pyrometers held their setpoint on a nitride-capped wafer while the wafer ran cold and its sheet resistance came out 26.31 % high. Its graded lot went from none of 25 in specification to all 25 once a film-stack table set the emissivity. The ion source's current reading stopped describing the source during a spark, and a gate that ignored it cut the overshoot from 14.75 to 0.29 %. The polisher's virtual metrology with an 18 nm bias reported 699.6 nm while the true film was 682.8, and the share in specification fell from 46 to 24 %. After a pad change the polisher that carried every estimate released production on a confident belief about a pad in the bin and landed 41.7 nm off target. A loop that holds its measurement has measured its sensor, and only the product says whether that was the right thing to hold.

12.2 Choosing a Method

The chapters suggest an order of questions for a tool that is new to the reader. Each question names the chapter where its answer decided the result, and Figure 12.2 draws the same questions as a chart.

1. *What does the tool measure, what does the product need, and when does each arrive?* Fill in a row of Table 1.1 before choosing anything. The answer separates loops that can correct this wafer from loops that can only correct a later one.
2. *Can the disturbance be seen or predicted before it reaches the product?* Feed it forward, from the surface ahead of the slit (Chapter 4), the step before (Chapter 6, Chapter 11), the charge passed (Chapter 10) or a known landing (Chapter 5), and leave the feedback a trim.
3. *How many runs pass before the reading returns?* Keep the weight times the gain ratio below the boundary of Equation (1.7), and match each reading to the recipe of its own lot, as Chapter 6 does. Count the age of a sampled analysis the same way (Chapter 10).
4. *Does the step end on an event?* End it on a signal from the wafer, an endpoint, a diameter or an inferred limit, rather than on a clock (Chapter 7, Chapter 2).
5. *Do several actuators act on one wafer?* Tune on the common move as well as the single zone, and model the coupling as a feedforward or a decoupler (Chapter 5, Chapter 9).
6. *Can the fast measurement be wrong while it looks right?* Validate it against the product, calibrate it for the film stack, and mark it invalid when its physical basis has gone (Chapter 9, Chapter 8).
7. *Is the fast actuator also the limited one, or does the plant grow during the run?* Return the actuator to mid-range with a slow one (Chapter 3), and schedule the gain on the variable that moves it (Chapter 2).

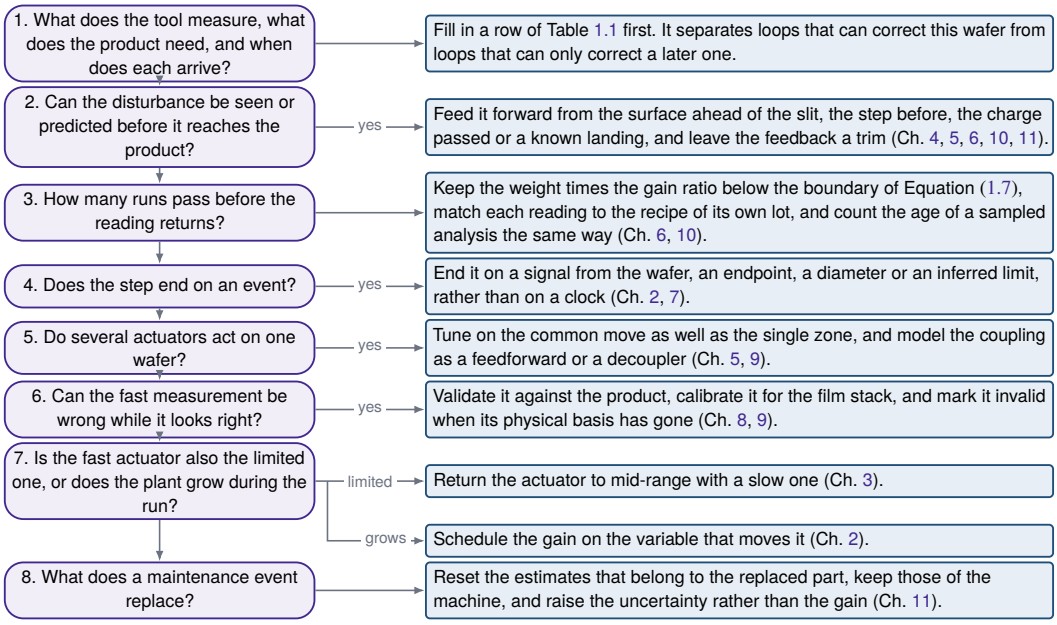


Figure 12.2 The eight questions of this section in order, with the list's answer and the chapters where it decided the result. Notice that the first three questions ask what is measured and when, and are answered before any loop is tuned.

8. *What does a maintenance event replace?* Reset the estimates that belong to the replaced part, keep those of the machine, and raise the uncertainty rather than the gain (Chapter 11).

A predictive controller paid where several actuators had to move together against limits. On the crystal puller it cut the diameter IAE of scenario 11 from 462.0 to 119.1 mm min with less heater travel, where loops retuned by the same rule reached 176.0 (Table 3.7). On the etch chamber it cut the time out of band after a recipe step from 37.08 to 3.54 s per wafer, but the loops retuned once reached 4.14 (Table 7.8). On the anneal's fast ramp it held the radial difference to 10.89 K against 17.27 for the decoupled loops, with the sheet resistance 3.34 % high, while the loops with a slip-aware ramp limit reached 6.29 K and −0.77 % (Table 9.7). On the polysilicon rod it saved time only by spending margin. With the band narrowed it harvested in 28.482 against 29.595 h with 4.4 °C left below the limit, and a retuned override reached 28.675 h with 9.1 (Table 2.6). Before reaching for prediction, try the retune, the supervisor and the feedforward. Volume 9 treats that choice in full.

12.3 Where Practice Is Heading

More data, used with process knowledge. A review of case studies in semiconductor manufacturing describes an industry that uses large data sets and analytics to improve fault detection and to support newer capabilities such as predictive maintenance. It finds that data quality is the most important factor for these solutions, and that subject matter expertise is often required to make them work on line [1]. In the editor's view the chapters of this volume show why. The anneal's pyrometers and the polisher's virtual metrology both produced clean, steady data that was wrong about the wafer, and only a model of the physics said which reading to distrust.

Virtual metrology between measurements. Virtual metrology estimates a product's quality from process data without measuring it, and a systematic review of 199 papers names machine control and

sampling decisions among its applications [2]. On an industrial plasma etch chamber, models updated over a moving window gave useful estimates of etch rate across several chamber maintenance events and many thousands of wafers [3]. That is the seasoning problem of Chapter 7, seen from the data. In the editor's view the estimate will increasingly sit inside the run-to-run loop, and Chapter 11 shows the cost when it does: a bias the loop cannot see becomes a bias in the product.

> ⧖ HISTORY **1991 to 2024: From a Proposal to the Fab's Default**
>
> The timeline of Chapter 1 placed the first run-to-run controller for chip fabrication in 1991. By 2008 a dissertation called run-to-run control the preferred technique for model-based process control, with the recipe adjusted lot by lot or wafer by wafer, and found most fabs combining it with fault detection and classification as advanced process control [4]. A 2018 review of metrology found that for some products more than half of the manufacturing steps can involve measurement [5]. A 2024 review of plasma etching named three-dimensional devices, critical dimensions approaching the atomic scale and new materials among its challenges, and digital twins of etch chambers built from hybrid models among its opportunities [6]. The loop of 1991 has become the fab's default, and the gaps of Section 12.5 are what it still lacks.

Sampling decided by risk. A review of inspection sampling classifies the techniques as static, adaptive and dynamic, and finds significant improvements from static toward dynamic sampling, which it judges more appropriate for modern plants [7]. Chapter 6 adds a control argument. With every lot measured, feedforward gave 1.050 nm of spread against 1.525 without it, so a sampling plan justified against the feedback loop alone undervalues each measurement.

Maintenance predicted, and models of the whole tool. Machine learning has been used to set maintenance decisions by the trade between unexpected breaks and unexploited lifetime, demonstrated on a semiconductor maintenance problem [8]. The pad of Chapter 11, the bath of Chapter 10 and the chamber walls of Chapter 7 are consumables of exactly that kind. Digital twins have been used for years in model-based process control and predictive maintenance, and a framework for them has been proposed with requirements of reuse, interoperability and autonomy over the twin's life [9]. In the editor's view each app of this volume is a small twin of a tool, and the record below shows the discipline a twin needs: it must keep the plant's truth on one side and the controller's knowledge on the other.

Anneals move to milliseconds. Flash lamp annealing heats on a sub-second time scale, away from equilibrium, and has been used in microelectronics mostly after ion implantation, to activate dopants, recrystallize amorphous layers and anneal out defects [10]. On a prototype flash system, a thermometer that compensates for emissivity read the wafer through a layer of flowing water that suppressed the flash, and its peak temperatures tracked the sheet resistance whatever the emissivity of the surface [11]. In the editor's view such a reading would answer the capped wafer of Chapter 9, which ran 31.4 K cold behind a steady trend.

Plating baths are read online. A portable microfluidic workstation reported in 2025 measured the additives of an acid copper bath online in small volumes, with a mean relative error below ten percent, in place of the manual stripping analysis of the laboratory [12]. A model-based approach to industrial electroplating takes the bath's composition from the data of the manufacturing execution and planning systems, because an exact analysis typically cannot be made within a plating company, and it was validated on two electrolytes [13]. In the editor's view both attack the 42 min age that set the trim limit of Chapter 10, one by shortening it and one by modeling across it.

Polishers learn from fab data. A typical fab generates millions of data points a year for process control and fault detection, and models trained on the data of a high-mix, high-volume fab predicted

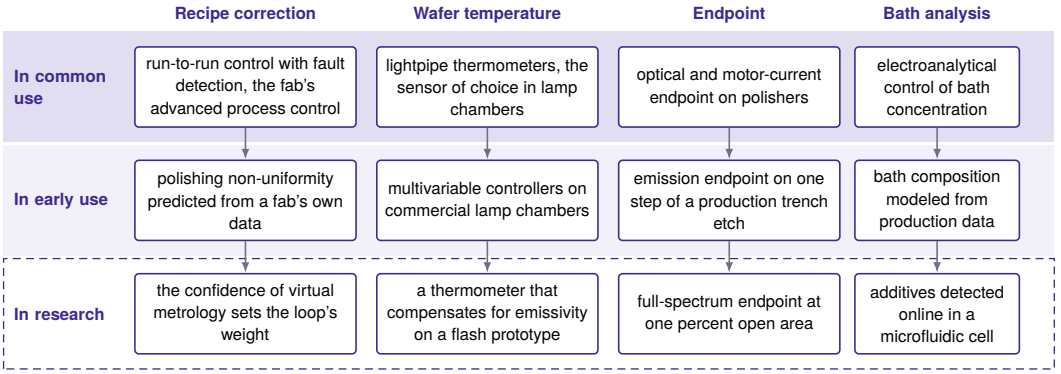


Figure 12.3 Developments in four threads. The top lane holds what the cited source calls standard or common practice, the middle lane what it reports in service at some sites or on plant data, and the bottom lane what it reports from simulation, laboratory work or as a need, in the editor's reading. Sources by thread: recipe correction [4, 14, 16], wafer temperature [11, 17, 18], endpoint [19–21] and bath analysis [12, 13, 22]. Notice that three of the four threads move the measurement closer to the wafer.

the non-uniformity and the spatially resolved removal rate of a polisher [14]. A 2025 review of machine learning in polishing finds that it can make control more adaptive and reduce the demand for metrology, and names data leakage, sparse labels of the target and hybrid modeling among its challenges [15]. In the editor's view the 18 nm bias that the virtual metrology of Chapter 11 carried unseen is the risk such a model must state. Figure 12.3 sets these developments on three lanes.

12.4 Engineering With AI Agents

An AI agent is a language model that works toward a goal in a loop. It plans, calls tools such as a compiler, a simulator or a browser, reads the results and continues [23]. In the editor's view, agents can now carry out much of the work that used to fill the hours of a control project. They derive a model from balances or fit one to plant data. They write the simulator and its tests, code the controller, tune it against stated criteria, script commissioning checks, write the documentation and watch running loops for degradation. The apps and chapters of this series were produced that way under the editor's direction, and the account that follows this section reports what the agents got wrong in this volume and how it was caught.

In the editor's view, this changes where an engineer's value lies. When a model, a controller and a report can each be produced in minutes, producing them is no longer the scarce skill. The work moves to five things.

Quality control. An agent's output arrives as a claim: this model fits, this tuning settles in four minutes, this code implements the equation. The engineer checks claims against the plant or the simulator. A reported number is measured again on a fresh run at the stated seed. The code is read for what it clamps, what it initializes and what it reads, because a display that shows the controller's model in place of the plant looks correct until the two differ. A tuning is tried on a disturbance the agent did not test. The most useful checks are independent of the author, made by a second agent or a person who has not seen the first answer, and they are checks that can fail.

Supervision. The engineer sets the scope of each task and decides what an agent may touch. Reading a historian and proposing a tuning is one level of access. Writing a setpoint, downloading a configuration or changing an interlock is another, and it passes through the plant's management of

change with a named approver. Supervision also fixes the order of work: a baseline before a change, one change at a time and a comparison that shows what moved.

Creativity. An agent answers the question it is given. Choosing the question remains the engineer's task: which loop is worth improving, whether a disturbance calls for feedforward, a cascade or a change to the process, which step test the operation can tolerate, and which comparison would show that a predictive controller earns its place over a retuned PID loop. The experiment that separates two explanations is designed by someone who can imagine both.

Process insight. A model contains what its author put in. The equipment does more. A valve sticks, the dead time of an analyzer grows when its sample flow drops, a thermowell lags, a heat exchanger fouls and an operator intervenes at the same hour each night. The engineer who knows the unit recognizes a result that is too good, asks what the model left out and knows which measurement to distrust first.

Accountability. Responsibility does not transfer to a tool. Functional safety standards assign the activities of the safety lifecycle to persons and organizations of stated competence, and they require an assessment that includes someone independent of the design team [24]. An agent may draft the hazard analysis or the test procedure. A person owns the decision to accept it, to start up and to keep running.

These five are the professional form of the habit introduced in Section 1.7: check before trusting, and know enough to check.

12.4.1 How This Volume Was Made

Each of the ten apps was built and tested by agents. An agent then wrote the paper on it, keeping verification notes that trace every number to a run, and a referee report was written on each paper. Six of the ten reports call themselves self-reviews. Defects went into a log for each app. For this book one agent per chapter reran the numbers its exercises use and checked the app's helpers, completion steps and figures. The editor set the scope of each round and the rules on what each agent could change, and statements that runs contradicted were left for the editor to decide.

The defect logs. Eight of the ten logs give each entry a severity, 129 entries in all, from 10 in the copper log to 26 in the bake plate's. Of these, 51 are wrong behavior or output of an app, 59 are documentation, 16 are cosmetic or display and 3 are other kinds. The scanner and etch logs number their items without a severity and are not added in. Five cases show the errors that matter for a volume about what a tool measures.

A timer that knew the etch rate. The etch app offered an operator aid, "Time the etch for this chamber". It set the main etch from the chamber's true etch rate, which no instrument on the tool measures. With a hidden etch-rate parameter raised by 20 % the aid shortened the etch from 86 to 73 s, as if it knew. A second aid set the actuators from hidden parameters of the chamber. In the anneal app the supervisor, the feedforward and the soak correction read the plant's own constants. Raising a parameter by 20 % moved the lamp commands by 18 to 150 W in one execution with no measurement changed. A truth audit found both by moving one hidden value at a time and asking whether the controller moved while the measurements did not. Both now compute from the declared model, and no shipped run moved. In a volume about endpoints, an aid that knew the rate had quietly removed the reason for an endpoint.

A field exposed one scan too long. The scanner ended a field when its scan coordinate reached the field length. At 400 mm/s the coordinate after 150 steps was 29.999999999999925 mm, so the test failed and every field ran one controller period longer. Each field received 1.33 % more dose, and the nominal mean linewidth was 61.665 nm instead of about 61.98. A dose bias of about 1.3 % had been noted earlier and not investigated. A consistency pass that re-measured every number traced it to the comparison. Every exposing run moved when it was fixed, and the paper was updated the same day.

Helpers pressed twice. Nine chapter records report a check of the app's helpers, and three report pressing a helper repeatedly. The ion source's "Set the bias current for the extraction setpoint" moved the bias by 0.06 A per milliampere of error where the source's sensitivity calls for about 0.0085. One press from the nominal state set 2.80 A and the extraction current went to 57.25 mA, a second set 1.77 A and 3.26 mA, and a third tripped the cathode. The polisher's "Set the polish time" scaled the time by the last wafer's removal, which had been made at the old time. The film came out at 699.92 nm after one press and at 754.33 after four. The bath's dosing aid, pressed six times, set 93.214 mL/h every time. A helper that answers once can still be a controller with the wrong gain when a person presses it again. Both helpers have since been corrected.

Grading against the learner's own number. The anneal challenge asks for every true center peak within 3 °C of 1050 °C. Its scoring judged the peak against the learner's Peak temperature setting instead. With that setting at 1048 °C and a soak of 1.10 s the lot was credited while its worst true peak was 4.23 K below 1050. The polisher's target thickness slider never reached its controller. With it at 720 nm the mean stayed at 699.92, yet three scenarios and the challenge judged against the slider, so lowering it to 685 credited a wear step on a fresh pad. Both were found by trying each rule with a value the learner can set. Both are the volume's lesson in another form: a grade on a setpoint measures the setpoint, not the wafer. Both rules now judge against the fixed target.

The book round. The ten chapter records list 147 rows of numeric verification, each a stated value or group of values set beside its reproduction. Eight report no drift. The polysilicon record found a table cell printed 1258.1 where the app gives 1258.03, and the ion source record found a −9.15 % species shift read at the first logged row, where the beam settles 14.4 % lower. Nine records report walking every completion step against its control's step and range. The book round's defect list holds eleven entries for this volume. Two apps label a direct-acting loop reverse acting: the puller's diameter and mid-ranging loops and the bath's bleed loop. The chapters name the action as Section 1.6.2 does, and the puller's and the bath's apps have since been corrected. Chapter 1 said that four of the etch chamber's loops held their setpoints to four decimal places. The etch agent found that three do, and the orchestrating agent corrected it and added a notation note from the bath chapter. The introduction's writer reported that a bake plate reference lacked an author, and the chapter agent's Crossref query showed the entry complete. Another session then changed the papers and apps of the polysilicon and puller chapters, and both chapters were synced again. Their app changes were text only, and no exercise answer moved. A later repair round changed six more of the volume's apps, and one answer of the etch chapter moved with them.

Every error above was caught by a check that could fail: a hidden value moved, a number traced to its run, a helper pressed twice, a rule tried with a setting the learner can change, a reference queried. None was caught by rereading. Four working rules follow. Give a controller and an aid only what the tool can measure. Press every aid twice. Grade against the product, never against a setting. Re-measure every number after every change.

12.5 Gaps in Semiconductor Process Control

Six gaps belong to this volume's topic, and the gaps common to agent-era practice follow them in Section 12.5.1. In the editor's view none is closed. A sentence on difficulty or progress that carries no

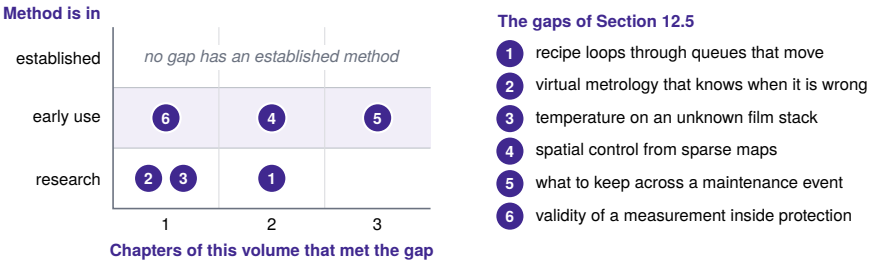


Figure 12.4 The six gaps of this section. Across is the number of chapters that each item names as evidence: two for gap 1 (Chapters 6 and 10), one for gap 2 (11), one for gap 3 (9), two for gap 4 (5 and 11), three for gap 5 (7, 10 and 11) and one for gap 6 (8). Up is the maturity of the method available, the editor's reading of the sources cited in each item. Notice that the gap most chapters met has a method in early use.

citation is the editor's view. Figure 12.4 places the gaps by the chapters that met them and by the maturity of the method available, and the same numbers mark the tools in the picture that opens this chapter.

1. *Recipe loops through queues that move.* The boundary of Chapter 6 assumes a fixed queue and a known gain ratio, and a gain believed at half its value took the spread to 7.712 nm. The age of the bath's analysis set the trim limit of Chapter 10 the same way. In a fab the queue is set by dispatch and changes from lot to lot. A design that states its margin against a queue and a gain that both vary is not routine. A 2008 dissertation reports that wafers wait in a backlog for metrology and are measured by priority, so the delay varies from run to run, mostly from 0 to 5 runs with a mean of 1 or 2 [4]. It is hard because the loop does not see the dispatch that sets its delay. Progress would be a weight chosen against the spread of the delay, with each reading's delay logged beside it.

2. *Virtual metrology that knows when it is wrong.* The polisher's estimate kept its loop on target while the product drifted 18 nm away (Chapter 11). An estimate needs a statement of its own error that a loop can act on, and a rule for when the loop should stop believing it. A 2018 study used the variance of a Gaussian-process prediction as its confidence and adjusted the weight of an EWMA run-to-run controller with it, on a polishing case study [16]. It is hard because a bias need not raise the spread that a model reports about itself.

3. *Temperature on an unknown film stack.* The anneal's pyrometers held a wafer 31.4 K cold without a sign on the trend, and a table or a calibration wafer corrected it only for stacks known in advance (Chapter 9). A 2005 study of a lamp-heated chamber reports that films on the measured side change the emissivity, that such wafers are screened out by hand, and asks for a systematic way to alarm the tool [25], and a lightpipe thermometer cannot give the true temperature unless the effective emissivity is known [26]. It is hard because the wrong emissivity leaves a steady, plausible trend. A measurement of the wafer's own emissivity, robust enough for every layer, would remove a whole class of silent error.

4. *Spatial control from sparse maps.* The bake plate's hundred offsets traded signature for roughness, and the polisher's eight chambers reach 33 radial points only through broad footprints (Chapter 5, Chapter 11). Methods that choose which spatial patterns a sparse map can correct, and leave the rest alone, are still tuned by hand. A 2004 study of a three-zone polisher read 60 points of film thickness, could not hold all 60 at their setpoints with three inputs, and controlled them through a singular value decomposition [27]. It is hard because the patterns that no actuator reaches still show in the map and pull the estimate.

5. *What to keep across a maintenance event.* The polisher's soft reset was right for a pad, the etch chamber's walls reseason after every clean, and the bath is dumped when bleeding cannot reopen its window (Chapter 11, Chapter 7, Chapter 10). A general method that splits an estimator's state by the part it describes and sets the uncertainty after the event is not available. The Open challenge box after this list takes the gap further.

6. *Validity of a measurement inside protection.* The ion source's gate decided in milliseconds whether a reading still described the source (Chapter 8). Each tool now carries its own such rules. A 1996 qualification study found an implanter highly susceptible to arcing, which can damage its hardware [28], and a 2021 system detects an implanter's faults in real time from tool data, because its wafers cannot be reworked [29]. It is hard because the rule must decide before the evidence is complete. A shared way to state, test and certify when a reading is valid, and what the loop does when it is not, would help every chapter of this volume.

OPEN CHALLENGE What Does a Maintenance Event Replace?

Gap 5 is the gap most chapters met. A run-by-run controller of 1995 gave maintenance its own rapid mode, which recovered an epitaxy process within three runs of a disturbance [30]. A 2008 dissertation reports a step in bias whenever a tool or chamber is maintained, and seemingly identical tools that process identical wafers differently with the number of lots since the last maintenance event [4]. In a production etch chamber the first one or two wafers of each lot differed considerably [20]. The event is hard because an estimator sees only the sum of its parts. Progress would be an estimator split by the part each state describes, with the maintenance record setting each part's uncertainty after the event.

12.5.1 Gaps Common to Agent-Era Practice

The final chapter of each volume lists gaps specific to its topic. Seven more are common to every topic in the series. In the editor's view none of them is closed.

1. *Verifying agent-written control code and models before they reach a plant.* Tests show that a controller behaves on the cases tested. For controllers learned from data, guarantees of stability and constraint satisfaction rest on assumptions about the model and its uncertainty, and open problems remain [31]. In the editor's view, code written by an agent has no comparable body of method yet. Practice needs a routine path from an agent's draft to code a plant can trust: independent derivation, tests of what the code reads and clamps, and trials with hardware in the loop.

2. *Trustworthy models from sparse or poor plant data.* Plant records are short, taken in closed loop, noisy and marked by unmeasured disturbances. A fitting routine returns a model from any record. What is missing is a dependable statement, delivered with the model, of where it can be trusted, and a method that says so when the data cannot identify a parameter.

3. *Safety cases and standards for agents in commissioning and operation.* Functional safety standards define a lifecycle of specified requirements, verification and assigned human roles [24]. In the editor's view, how an agent that writes configuration, runs commissioning checks or advises an operator is qualified, limited and audited under them is not yet settled.

4. *Security of agent access to control systems.* An agent with credentials to a control system is a new path into it. Industrial security standards state the requirements a control system must meet at each security level [32]. An agent that reads documents, messages or web pages can also be steered by text placed there. In the editor's view, least privilege, read-only access by default and human approval of every write are a starting point and not a solution.

5. *Benchmarks and reproducibility.* An agent asked twice may deliver two different controllers, and a result reported without its plant, seed and scoring cannot be checked. Control needs shared

benchmark plants with fixed disturbances and published measures, on which agent-built and conventional designs are compared like for like. The apps of this series, with a default seed and an exported record of every run, are a start.

6. *Transfer from simulation to hardware.* A controller developed on a simulator inherits what the simulator omits. The robotics literature documents the gap between simulated and real performance and the methods used to narrow it, such as randomizing the simulated plant and adapting on the real one [33]. Process plants add slow dynamics, rare upsets and little freedom to experiment, so each transfer still needs a staged plan from the simulator to a bench rig to the unit.

7. *Keeping engineers' judgment sharp when agents do the routine work.* Automation removes the routine work through which skill is built and leaves the person the hard cases, an irony described for process operators in 1983 [34]. The same now applies to engineers. An engineer who has never tuned a loop by hand or closed a balance on paper is poorly placed to review an agent that has. Deliberate practice, of the kind the exercises in this book provide, has to be designed into education and into work.

KEY POINTS

- What can be seen before it reaches the wafer is fed forward, and the feedback behind it is left a trim.
- A late reading sets the usable gain of a recipe loop, and so does how the estimator matches each reading with its lot.
- A step that ends on a signal from the wafer absorbs a drift that no loop on the chamber can see.
- Zones on one wafer move together and need a model of how they share it.
- A loop that holds its measurement has measured its sensor. Only the product says whether that was the right thing to hold.

References

[1] J. Moyne and J. Iskandar, "Big data analytics for smart manufacturing: Case studies in semiconductor manufacturing," *Processes*, vol. 5, no. 3, p. 39, 2017.

[2] P.-A. Dreyfus, F. Psarommatis, G. May, and D. Kiritsis, "Virtual metrology as an approach for product quality estimation in Industry 4.0: A systematic review and integrative conceptual framework," *International Journal of Production Research*, vol. 60, no. 2, pp. 742–765, 2022.

[3] S. A. Lynn, J. Ringwood, and N. MacGearailt, "Global and local virtual metrology models for a plasma etch process," *IEEE Transactions on Semiconductor Manufacturing*, vol. 25, no. 1, pp. 94–103, 2012.

[4] A. V. Prabhu, "Performance monitoring of run-to-run control systems used in semiconductor manufacturing," Ph.D. dissertation, The University of Texas at Austin, 2008. [Online]. Available: http://hdl.handle.net/2152/17721

[5] N. G. Orji, M. Badaroglu, B. M. Barnes, C. Beitia, B. D. Bunday, U. Celano, R. J. Kline, M. Neisser, Y. Obeng, and A. E. Vladar, "Metrology for the next generation of semiconductor devices," *Nature Electronics*, vol. 1, no. 10, pp. 532–547, 2018.

[6] G. S. Oehrlein, S. M. Brandstadter, R. L. Bruce, J. P. Chang, J. C. DeMott, V. M. Donnelly *et al.*, "Future of plasma etching for microelectronics: Challenges and opportunities," *Journal of Vacuum Science & Technology B*, vol. 42, no. 4, p. 041501, 2024.

[7] J. Nduhura-Munga, G. Rodriguez-Verjan, S. Dauzère-Pérès, C. Yugma, P. Vialletelle, and J. Pinaton, "A literature review on sampling techniques in semiconductor manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 26, no. 2, pp. 188–195, 2013.

[8] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi, "Machine learning for predictive maintenance: A multiple classifier approach," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 3, pp. 812–820, 2015.

[9] J. Moyne, Y. Qamsane, E. C. Balta, I. Kovalenko, J. Faris, K. Barton, and D. M. Tilbury, "A requirements driven digital twin framework: Specification and opportunities," *IEEE Access*, vol. 8, pp. 107 781–107 801, 2020.

[10] L. Rebohle, S. Prucnal, Y. Berencén, V. Begeza, and S. Zhou, "A snapshot review on flash lamp annealing of semiconductor materials," *MRS Advances*, vol. 7, no. 36, pp. 1301–1309, 2022.

[11] Y. Yamada, T. Aoyama, H. Chino, K. Hiraka, J. Ishii, S. Kadoya, S. Kato, H. Kiyama, H. Kondo, T. Kuroiwa, K. Matsuo, T. Owada, T. Shimizu, and T. Yokomori, "In situ Si wafer surface temperature measurement during flash lamp annealing," *Japanese Journal of Applied Physics*, vol. 49, no. 4S, p. 04DA20, 2010.

[12] Y. Zhao, J.-X. Zeng, J.-Q. Yang, T. Song, R. Hu, J.-J. Su, B. Zhang, F.-Z. Yang, D. Zhan, and L. Han, "On-line detection of additive concentrations in acidic copper plating solution for metal interconnection by an electrochemical microfluidic workstation," *Industrial Chemistry & Materials*, vol. 3, no. 5, pp. 607–617, 2025.

[13] A. Leiden, S. Kölle, S. Thiede, K. Schmid, M. Metzner, and C. Herrmann, "Model-based analysis, control and dosing of electroplating electrolytes," *The International Journal of Advanced Manufacturing Technology*, vol. 111, no. 5-6, pp. 1751–1766, 2020.

[14] M. Breidung, T. Rothe, A. Lauff, P. Thieme, J. Langer, M. Günther, and H. Kuhn, "Process data-driven machine learning for non-uniformity prediction and virtual metrology in chemical mechanical planarization," *Journal of Intelligent Manufacturing*, 2025.

[15] G. Winkler, T. Rothe, M. A. Sayyed, L. Jäckel, J. Langer, H. Kuhn, and M. Stoll, "Machine learning in chemical-mechanical planarization: A comprehensive review of trends, applications, and challenges," *Advanced Engineering Informatics*, vol. 68, p. 103663, 2025.

[16] J. Wan and S. McLoone, "Gaussian process regression for virtual metrology-enabled run-to-run control in semiconductor manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 31, no. 1, pp. 12–21, 2018.

[17] C. W. Meyer, D. P. DeWitt, K. G. Kreider, F. J. Lovas, and B. K. Tsai, "ITS-90 calibration of radiation thermometers for RTP using wire/thin-film thermocouples on a wafer," *AIP Conference Proceedings*, vol. 550, pp. 254–258, 2001.

[18] A. Emami-Naeini, J. L. Ebert, D. de Roover, R. L. Kosut, M. Dettori, L. M. L. Porter, and S. Ghosal, "Modeling and control of distributed thermal systems," *IEEE Transactions on Control Systems Technology*, vol. 11, no. 5, pp. 668–683, 2003.

[19] H. Jeong, H. Kim, S. Lee, and D. Dornfeld, "Multi-sensor monitoring system in chemical mechanical planarization (CMP) for correlations with process issues," *CIRP Annals*, vol. 55, no. 1, pp. 325–328, 2006.

[20] S. Lynn, "Virtual metrology for plasma etch processes," Ph.D. dissertation, National University of Ireland Maynooth, 2011. [Online]. Available: https://mural.maynoothuniversity.ie/id/eprint/2657/

[21] S. Lee, H. Choi, J. Kim, and H. Chae, "Spectral clustering algorithm for real-time endpoint detection of silicon nitride plasma etching," *Plasma Processes and Polymers*, vol. 20, no. 6, p. e2200238, 2023.

[22] R. Carpio and A. Jaworski, "Review—management of copper damascene plating," *Journal of The Electrochemical Society*, vol. 166, no. 1, pp. D3072–D3096, 2019.

[23] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J. Wen, "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, no. 6, p. 186345, 2024.

[24] International Electrotechnical Commission, "Functional safety – Safety instrumented systems for the process industry sector – Part 1: Framework, definitions, system, hardware and application programming requirements," IEC 61511-1:2016, Edition 2.0, 2016.

[25] W. Cho, "Temperature control and modeling of rapid thermal processing," Ph.D. dissertation, The University of Texas at Austin, Austin, TX, 2005. [Online]. Available: https://hdl.handle.net/2152/1527

[26] Y. Qu, "Silicon wafer surface temperature measurement using light-pipe radiation thermometers in rapid thermal processing systems," Ph.D. dissertation, The University of Texas at Austin, Austin, TX, 2006. [Online]. Available: https://hdl.handle.net/2152/2794

[27] S.-J. Shiu, C.-C. Yu, and S.-H. Shen, "Multivariable control of multizone chemical mechanical polishing," *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 22, no. 4, pp. 1679–1687, 2004.

[28] S. K. Joung, "Qualification of a medium current ion implantation system in a semiconductor production environment," S.M. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1996. [Online]. Available: https://hdl.handle.net/1721.1/40220

[29] S. C. Horng and Y. L. Lin, "Fuzzy tree classification system for fault diagnosis on ion implanter," *Journal of Physics: Conference Series*, vol. 2020, no. 1, p. 012017, 2021.

[30] E. Sachs, A. Hu, and A. Ingolfsson, "Run by run process control: combining SPC and feedback control," *IEEE Transactions on Semiconductor Manufacturing*, vol. 8, no. 1, pp. 26–43, 1995.

[31] L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig, "Safe learning in robotics: From learning-based control to safe reinforcement learning," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 411–444, 2022.

[32] International Electrotechnical Commission, "Industrial communication networks – Network and system security – Part 3-3: System security requirements and security levels," IEC 62443-3-3:2013, Edition 1.0, 2013.

[33] W. Zhao, J. P. Queralta, and T. Westerlund, "Sim-to-real transfer in deep reinforcement learning for robotics: A survey," in *Proc. 2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2020, pp. 737–744.

[34] L. Bainbridge, "Ironies of automation," *Automatica*, vol. 19, no. 6, pp. 775–779, 1983.

12.6 Exercises

12.1 Discussion · 8 min

An agent proposes to raise the etch loop's EWMA weight on the line of Chapter 6 from 0.30 to 0.60, because "the line drifts and the loop is slow". Its evidence is a simulation with no metrology queue in which the corrected line returns to target twice as fast. Using Chapter 6 and Section 1.4.4, explain why the evidence does not support the change, and say what you would do instead.

12.2 Discussion · 7 min

An agent reports that it passed the graded lot of Chapter 9 with the suggested structure, a Peak temperature of 1048 °C and a soak of 1.10 s. The Lot target tile reads met, 25 of 25 in specification. Using Chapter 9 and Section 12.4.1, say what the report does and does not show, and what you would check before accepting it.

12.3 Discussion · 8 min

A new single-wafer deposition tool has five heater zones under the wafer. Film thickness is measured after deposition on one wafer in five, and the reading returns three wafers late. The incoming wafer's prior film is measured on every wafer at the step before. The chamber walls are cleaned every few days and the rate drifts after each clean. Use the questions of Section 12.2 to propose a control structure, and name the chapter that supports each element.

12.4 Discussion · 7 min

To avoid qualification wafers after every pad change, an agent proposes that the polisher of Chapter 11 carry every estimate across the change. It argues that the estimator is confident and that the release rule will hold production if anything is wrong. Evaluate the proposal using Chapter 11, and compare it with how Chapter 7 and Chapter 6 treat a wet clean.

12.5 Discussion · 8 min

A team must decide whether to add a predictive controller to the bake plate of Chapter 5 or to the plating bath of Chapter 10. Using the four comparisons of Section 12.2 (crystal puller, etch chamber, anneal, polysilicon rod), state where prediction paid in this volume and where it did not, and argue for one of the two candidates or for neither.

12.6 Discussion · 7 min

An agent has written three new operator aids for a fab's tools and reports that each sets its tool correctly when pressed. Using the cases of Section 12.4.1, write a short acceptance test for agent-written aids and say which failure of this volume each step would have caught.

Answers to Selected Exercises

Short answers are given for the odd-numbered exercises. Numeric answers from an app are for its default seed and settings unless the exercise says otherwise.

Chapter 1. Introduction

1.1 (a) 0.855 nm, 6.9 % more than with no control. (b) 0.160 nm.

1.3 (a) 1.000 at $m = 1$ and 0.445 at $m = 3$. (b) No: $\xi = 1.25$ and $K = 0.50 > 0.445$.

1.5 (d). With the gain known, regressing on the recipe that made the lot removes the delayed term, and the loop is stable for any weight below 2.

1.7 611.27 °C, 11.27 K above the setpoint.

1.9 (a) 0.500 nm. (b) $g_* = 0.900$, 0.474 nm. (c) 1.118 nm.

Chapter 2. Siemens Polysilicon Reactor Gain Scheduling

2.1 (d). Under voltage regulation $\partial P/\partial \bar{T} = +PB/\bar{T}^2$, which is 2.31 times the loss slope at 60 mm.

2.3 (b). The model carries boron only into the purity of the deposit, which no instrument measures.

2.5 (a) 461.7 W/K. (b) 1067 W/K, 2.31 times the loss slope, so only current regulation is stable. (c) 59.4 s.

2.7 (a) 1053.2 °C, IAE 65.0 °C h. (b) IAE 5.9 °C h, a factor of 11, overshoot 15.6 %, multiplier 51.2. (c) The tile reads the noisy pyrometer, whose 2 °C noise is a tenth of the step.

2.9 (a) 29.68 h and 1348.8 °C, 28.86 h and 1374.5 °C, 30.62 h and 1323.1 °C. (b) Bands 80 and 120 pass. Margin costs 0.032 to 0.036 h of batch per degree.

2.11 (a) About \$1114 per batch, \$5.22 per kilogram. (b) 2.94 MV A, 6.5 times the largest power, so the supply is built with several voltage ranges.

Chapter 3. Czochralski Crystal Puller Mid-Ranging

3.1 (c). With no derivative action the double integrator has a negative phase margin, and the meniscus oscillation grows until it floods.

3.3 (a). Equation (3.14): K_c scales with V_p/V_0 and every controller time with V_0/V_p.

3.5 (a) 0.1647 kg/min, 4.94 kW. (b) $T_c = 1582.8\,^{\circ}\mathrm{C}$, $T_h = 1727.9\,^{\circ}\mathrm{C}$, $P_a = 133.8\,\mathrm{kW}$. (c) 15.6 kW above the 118.26 kW design power, about 9.6 kW of it the melt surface loss.

3.7 (a) 2.64 h (158.4 min); the meniscus floods. (b) 309.8 mm at 1.70 h. (c) 0.214 mm/min, so the loss is through the meniscus and not the structure limit.

3.9 (a) 1.89 h with 19.0 kg prime; 11.25 and 3.75. (b) 8.43 h, 88.0 kg, 0.086 mm/min. (c) 23.0° of phase margin.

3.11 (a) \$79,200 against \$66,116, about \$13,100 per boule, plus an eight-hour furnace run, the polysilicon and the electricity. (b) The mid-ranged heater has absorbed a disturbance that no instrument reports, and its drift is the record of it.

Chapter 4. Wafer Scanner Preview Feedforward

4.1 (b). The needed lead is $v\tau_g$, which grows from 1.70 to 2.97 mm while the preview stays where it was.

4.3 (d). The trim's proportional term subtracts $K_{c,f} d_{trail}$ times the slope, 0.60 mm of lead at the shipped trim.

4.5 (a) 90 and 51.4. (b) 1.58 % and 2.09 %, ratio 1.32. (c) 0.44 mJ and 0.88 mJ, inside the limit.

4.7 (a) 119.0 nm. (b) 62.9 nm and 41.7 nm. (c) 13.8 nm, so the best the sliders allow from the sensor behind is 3.0 times worse than the sensor ahead on the same servo.

4.9 (a) 5.12 ms, 2.46 mm, then 2.46 ms, 1.18 mm. (b) 24.2 nm, target not met at 69.3 nm. (c) 11.1 nm, target met with a worst sample of 34.1 nm.

4.11 (a) 1.33 Hz and 57.1 Hz, then 2.67 and 163 μm Hz^2, a ratio of 61. (b) The topography, specified as site flatness over one field.

Chapter 5. Hundred-Zone Bake Plate Decentralized Control

5.1 (a). The ratio is $(U_a + 8k_x)/U_a$, so halving k_x takes it from 7.86 to 4.43.

5.3 (c). Each loop reads the pulse as its own error, its trim falls, and the sag that follows costs more dose than the pulse added.

5.5 (a) 98.42 °C. (b) 30.800 W, 32.537 W, 34.274 W; 3156.4 W, which is 76.4 W more. (c) 31.6 s and 5.16 s, a ratio of about six.

5.7 (a) 62.11 nm and 0.27 nm; 65.59 and 0.02 nm; 57.07 and 0.80 nm. (b) 119.29 °C, 9.29 °C above setpoint. (c) The linewidth is a dose, and the dose lost in the dip is small. The pulse is sized to the dose of the calibrated bake.

5.9 (a) 0.87 nm and 0.22 per zone; 0.76 nm and 0.16 filtered. (b) The filter removes the high-wavenumber content of each update, which is mostly measurement noise that the plate could deliver only at 7.86 times the power per degree.

5.11 (a) The three neighbor loops hold their own zones within 0.003 °C, but nothing regulates the silicon above them. (b) The saturation flag, with the faulted zone commanded at 120 W and delivering 14.40 W.

Chapter 6. Lithography and Etch Run-to-Run Control

6.1 (a). The wall condition changes the etch bias by $\beta\theta$, and no sensor on the line reads it.

6.3 (c). The developed linewidth of a lot is read before that lot is etched, so the trim cancels the lithography term on the same lot.

6.5 (a) 1.200, 1.200, 1.200, 0.840, 0.480, 0.120, −0.132 nm; 10.720 s. (b) 0.840, 0.588, 0.412 nm. (c) Only the newest-reading estimator corrects the same error again on lots 10 and 11.

6.7 (a) 8.00 nm, 15 %; 1.27 nm, 82 %; 1.55 nm, 82 %. (b) Only the first run, at 0.95 against 0.618, is outside its boundary. (c) The matched lot is stable at 0.95 but still writes most of each noisy reading into the next recipe.

6.9 (a) 1.85 nm, 67 %; 0.85 nm, 100 %; 0.64 nm, 100 %. (b) 0.570 nm. The third run agrees within the scatter of a 40-lot estimate. (c) Feedback at 0.30 adds back the metrology noise it chases.

6.11 (a) Recovery from a disturbance takes three times as long in production time, and two lots in three are etched with no knowledge of their own incoming linewidth. (b) The after-develop reading, since only it carries the feedforward.

Chapter 7. Plasma Etch Recipe and Endpoint Detection

7.1 (d). The coverage θ_w climbs from 0.078 to 0.264 by wafer 4, and the etch rate rises with it while every loop reads its setpoint.

7.3 (b). It compares the emission, filtered over 0.30 s, with 0.55 and moves the sequencer from main etch to over-etch.

7.5 (a) $\eta = 0.484$, $s = 0.150$. (b) −3.50 °C/Torr. (c) 0.574 W/K, 104.5 s and 0.843 °C/W, about seven times the model's values.

7.7 (a) At 5 % wafers 1 and 2 clear only 0.987 of the wafer with 1.15 nm of loss and are scrapped. At 30 % wafers 3 to 5 clear and lose 6.16 nm. At 75 % wafer 6 clears and loses 15.18 nm, past the limit. (b) 0.23 nm/s; about 59 %.

7.9 (a) 7 / 7, which meets the target; 5 / 7, which does not. (b) Wafers 5 and 6, not cleared. (c) The endpoint ends each main etch when that wafer's film has cleared, whatever the rate was.

7.11 (a) Three wafers, about 331 s of tool time. (b) At best wafers 2 onward, since no reading exists before wafer 1 is etched. The endpoint ends each etch on the wafer itself and passes all eight.

Chapter 8. Ion Source Cascade and Fault Recovery

8.1 (c). The inner loop holds extraction current with the bias supply, the fastest electrical handle on the cathode.

8.3 (a). The detector of Equation (8.22) needs a surge and a collapse together, and a gas loss produces neither.

8.5 (a) 0.23 ms and 0.55 ms. (b) 0.36 kV at 2.66 ms, about 2.43 ms of coincidence, so a 3 ms window latches no spark. (c) 8.0 J, 0.32 J, 32.64 J.

8.7 (a) 0.0746, 14.8 %, 50 ms and 0.0083, 0.3 %, 51 ms. (b) 77 ms, REPAINT at 3.64 s against 3.26 s. (c) Only the quench dwell delays the beam's return.

8.9 (a) 0.946 mA, 0 of 5, J chase 0.0000. (b) 0.117 mA, 2 of 5. (c) 0.946 mA, J chase 0.3139 A s.

8.11 Extraction is plasma limited, 33.20 mA against a 55.45 mA space-charge limit, so voltage buys no current and changes the implant energy, and the filament acts through two thermal lags.

Chapter 9. Rapid Thermal Processing Ramp Feedforward

9.1 (b). A gain of 4/s on the heating rate dwarfs the radiative rate of 0.15 to 0.42/s.

9.3 (d). Without the feedforward the whole wafer lags the ramp and overshoots the peak by about 30 K.

9.5 (a) 1.8 s, 31.5 kW. (b) 62.5 K, close to the measured 65.2 K. (c) 2.37 s, longer than the ramp.

9.7 (a) 355.3 Ω/sq, a lag of about 64.9 K. (b) 404.7 Ω/sq, a lag of about 2.6 K. (c) The integral overshoots the peak by about 30 K, where most of the budget accrues.

9.9 (a) 0 in spec, 15.34 %, 27.8 K, 23 slips. (b) 25 in spec, 1.23 %, 2.3 K, no slip, 3.4 kW h. (c) The film-stack table, because wafers 10 to 15 are nitride capped.

9.11 The table costs nothing but trusts the declared stack; a calibration wafer costs a wafer and chamber time but measures the stack, and goes stale as the chamber or the stack changes.

Chapter 10. Copper Plating Bath Dosed by Amp-Hours

10.1 (b). The analysis reports 40 min after a sample drawn through a 2 min line, so the 4:40 result describes the bath of 3:58.

10.3 (d). Its output rises with the measurement, so it is direct acting, as Table 10.3 records ($\gamma = -1$).

10.5 (a) 45.4 mL/h and 12.0 mL/h; at $C = S_V$ the bleed takes out what the make-up brings in; 57.4 mL/h and 20.9 h. (b) 55.16 mL/h and 2.27 mL/h. (c) 9.8 % and 3.6 %; $r_c = 3.67$ with a lower edge of 3.35, so a ratio of 3.0 is below the window.

10.7 (a) 0 voids, worst margin 0.23; 57.4 mL/h at 4.0 h and 101.8 mL/h at 6.0 h. (b) 97 voids, -0.06; 5.96 mL/L, then 5.70 mL/L; first void at 8.1 h. (c) The first sample after the lot is drawn at 6:00, and the trim adds about 10 mL/h a report.

10.9 (a) 226 voids, -0.37, 12.0 L, not met. (b) 0 voids, 0.12, 57.0 L, met; $B = 24.55\,\mathrm{mL/L}$; exit mean 23.8 mL/L. (c) 0, 0.05, 38.7 L, met; exit mean 25.1 mL/L; about 9.2 L per mL/L.

10.11 (a) The actuator is one-sided, and the leveler falls only by consumption, breakdown and bleed, to 3.41 mL/L by 24 h. (b) The dump at 3.1 h saves 195 voids at a cost of 34 wafers and 115.8 L more make-up. A pump interlock or a dose totalizer would have stopped it.

Chapter 11. Planarization Profile Control Over a Pad Life

11.1 (a). A thick wafer polished exactly as predicted reads as a perfect polish, so the residual is zero.

11.3 (c). The footprint width grows with the band width (Equation (11.1)).

11.5 (a) 0.82, 90 nm high. (b) $W = 0.2647$; wafer 70. (c) $A^* = 0.9083$ and 0.9474; 236.6 wafers, so the pad ends at wafer 237.

11.7 (a) 1, 0.0070, 2.5 nm. (b) 1, 0.0100, 41.7 nm. (c) 6, 0.0500, 59.4 nm.

11.9 (a) 0.14 to 4.81 nm; 19.00 nm; 18.38 nm. (b) 9.03 nm; 20.48 nm. (c) 15 nm, 100 %; at 16, 99 %.

11.11 (a) The bias enters the residual as a process error, so the recipe removes it and the surrogate then reads target. (b) 5.7 times the noise, above the threshold of 4; sample more, never less.

Chapter 12. Summary and Outlook

12.1 A good answer notes that the test left out the two-lot queue, which sets the boundary at 0.618 for the newest-reading estimator, so 0.60 sits at the edge and a trim rate believed at half its value doubles the loop gain. On lithography noise the weight of 0.60 gave a spread of 2.663 nm against 1.852 at 0.30 and 1.476 with no control. Feedback cannot remove white lot-to-lot noise. Feedforward from the developed line gave 0.850 nm at the unchanged weight. It asks for a run with the queue, a gain error and the measured spread.

12.3 A good answer fills in the row first: thermocouples fast, thickness late and sampled, incoming film known. It feeds forward the incoming film (Chapter 6, Chapter 11), runs a run-to-run loop on thickness with a weight inside the boundary for $m = 3$ and a matched-wafer estimator (Chapter 6), tunes the zones on the common move and models their coupling (Chapter 5, Chapter 9), and resets only the wall estimate after a clean with raised uncertainty (Chapter 11, Chapter 7). It names an endpoint or virtual metrology as an option if the rate cannot be held.

12.5 A good answer summarizes the four. The puller reached 119.1 against 176.0 retuned and 462.0 shipped, with less heater travel. The etch chamber reached 3.54 against 4.14 retuned. The anneal held 10.89 K against 17.27 for decoupled loops but with the sheet resistance 3.34 % high, where a slip-aware ramp limit gave 6.29 K and -0.77 %. The rod was faster only by spending margin. Prediction paid where actuators interact under limits. The plate's coupling is already handled by a static model and the landing by a pulse. The bath's limit is a two-hour analysis, which a model cannot shorten. A good answer argues for neither until a retune and a feedforward have been tried, or names the test that would decide.

Glossary

Each term is given as this volume uses it, in one plain sentence, followed by the chapter where it is first used in earnest. The two indexes give every page.

advanced process control (APC) The fab's control system above the tools, which combines run-to-run control with fault detection and classification. *Chapter 1*.

amp-hour counter A meter that adds up the electric charge a plating rectifier has passed, from which the additives used up are estimated. *Chapter 10*.

analyzer An instrument that takes a sample and reports a composition some time later; on the plating bath it measures the additives. *Chapter 10*.

back-calculation A way to stop windup: when a controller's output is limited, its integral is pulled toward the output actually applied. *Chapter 1*.

bake plate A heated plate, divided into zones, on which an exposed wafer is baked so that the chemistry of the resist finishes. *Chapter 5*.

Bayesian estimation Estimating unknown parameters by combining what was believed beforehand with new measurements, each weighted by its uncertainty. *Chapter 11*.

bleed and feed Draining a little plating bath all the time and replacing it with fresh solution, so that breakdown products cannot build up. *Chapter 10*.

breakdown products Compounds that the plating additives turn into as they are used, which collect in the bath and narrow the fill window. *Chapter 10*.

bright-ring camera A camera that watches the ring of reflected light where the melt meets the growing crystal, to measure the crystal's diameter. *Chapter 3*.

carrier head The part of a polisher that holds the wafer face down on the pad and presses it through several pressure chambers. *Chapter 11*.

cascade control Two loops in series, in which the outer loop sets the setpoint of a faster inner loop. *Chapter 3*.

CD-SEM A scanning electron microscope built to measure linewidths on wafers quickly and without damaging the feature. *Chapter 6*.

chamber seasoning The slow change of an etch chamber's walls as deposits build up after a cleaning, which moves the etch rate from wafer to wafer. *Chapter 6*.

chemical mechanical planarization (CMP) Polishing a wafer flat with a pad and a slurry that both wears and chemically softens the surface. *Chapter 11*.

chemically amplified resist A resist in which light releases an acid that sets off many reactions during the bake, so that little light is needed. *Chapter 5*.

Child-Langmuir law The law by which the current an extraction gap can draw from a plasma grows with the voltage to the power three halves. *Chapter 8*.

clearing transition The interval in which a film disappears from different parts of the wafer at different times, because the etch rate varies. *Chapter 7*.

copper electroplating Depositing copper on a wafer from a solution by passing an electric current, used to fill the trenches of chip wiring. *Chapter 10*.

critical dimension (CD) The width of the smallest feature a step must print or etch, such as a transistor's gate line. *Chapter 4*.

cyclic voltammetric stripping (CVS) An analysis that plates and strips copper on a rotating electrode in a sample; the stripping charge shows the additives' strength. *Chapter 10*.

Czochralski crystal growth Pulling a single crystal slowly upward from a melt, starting from a small seed that sets its orientation. *Chapter 3*.

damascene copper Wiring made by etching trenches in an insulator, filling them with plated copper and polishing the excess away. *Chapter 10*.

decentralized control Control by many single loops, each reading one sensor and moving one actuator, with no loop told what the others do. *Chapter 1*.

defect band The narrow range of growth rate in which a crystal freezes with few enough point defects to make device wafers. *Chapter 3*.

dopant activation Heating an implanted wafer so that the dopant atoms take places in the crystal and begin to carry current. *Chapter 9*.

dose ledger A running record, strip by strip, of the ion dose each part of the wafer has received and should have received. *Chapter 8*.

double EWMA A run-to-run estimator with a second average for the slope, which in theory removes the offset that a steady drift leaves. *Chapter 1*.

emissivity How strongly a surface radiates compared with a perfect radiator at the same temperature; a pyrometer must be told it. *Chapter 1*.

endpoint detection Ending a step when a signal shows that the film has cleared, instead of when a

timer runs out. *Chapter 1*.

etch recipe The setpoints of pressure, power and gas flow for each step of an etch, developed by experiment for each product. *Chapter 7*.

EWMA controller A run-to-run controller that estimates the disturbance as an exponentially weighted moving average of what each reading implies. *Chapter 1*.

exposure dose The light energy per unit area that the resist receives, which sets how wide the printed line comes out. *Chapter 4*.

Faraday cup A metal cup that catches ions and measures their current, used to read the beam beside the wafer. *Chapter 8*.

fault detection and classification (FDC) Watching a tool's sensor data to recognize an abnormal run and to name its likely cause. *Chapter 1*.

feedforward Correcting for a disturbance from a measurement of it, before it shows in the controlled result. *Chapter 1*.

fill window The range of additive and breakdown-product levels within which plating fills trenches from the bottom up without voids. *Chapter 10*.

first-wafer effect The different etch of the first wafers after a chamber has been cleaned, before its walls settle. *Chapter 7*.

Freeman source An ion source with a straight hot filament beside the extraction slit, long the most widely used source for doping. *Chapter 8*.

gain mismatch The ratio of a process's true gain to the gain its controller assumes, which multiplies the loop gain. *Chapter 6*.

gain scheduling Changing a controller's tuning as a measured variable changes the process, here the diameter of a growing rod. *Chapter 2*.

in-line metrology Measurements of the wafer made in the fab between process steps, usually on a sample of the wafers. *Chapter 1*.

inferential control Control of a quantity that is not measured, through an estimate of it computed from quantities that are. *Chapter 1*.

influence matrix A table of how strongly each pressure chamber of the carrier head changes the removal at each radius of the wafer. *Chapter 11*.

integral hold Freezing a loop's integral while a planned feedforward pulse runs, so that feedback does not undo the pulse. *Chapter 5*.

ion implantation Placing dopant atoms in a wafer by firing them in as an accelerated beam of ions sorted by mass. *Chapter 8*.

ion source The part of an implanter that turns a gas into a plasma and extracts ions from it as a beam. *Chapter 8*.

Kalman filter An estimator that weighs each new measurement against a model by their uncertainties; for a drifting disturbance its steady gain is the best EWMA weight. *Chapter 1*.

level sensor An optical sensor that reads the height of the wafer's surface ahead of the exposure slit. *Chapter 4*.

lithography scanner The tool that prints a pattern into resist by projecting light through a mask, field by field. *Chapter 4*.

lot A group of wafers, often 25, that travel between tools and are processed together. *Chapter 1*.

matched-lot estimator A run-to-run estimator that pairs each late reading with the recipe of its own lot, which lifts the delay's limit on the gain. *Chapter 1*.

meniscus The curved surface of liquid between the melt and the bottom of the growing crystal. *Chapter 3*.

metrology Measurement of the wafer itself, such as a linewidth, a thickness or a resistance, as opposed to a tool's own sensors. *Chapter 1*.

metrology delay The number of runs that pass between a run and the arrival of its measurement. *Chapter 1*.

metrology sampling Measuring only some lots or wafers, which lengthens the time between corrections. *Chapter 1*.

mid-ranging control A structure in which a slow actuator returns a fast but limited one toward the middle of its range. *Chapter 3*.

model predictive control (MPC) A controller that uses a model to plan several moves ahead against limits, applies the first and plans again. *Chapter 1*.

multizone thermal control Control of one continuous plate or wafer by many heaters and sensors that act on each other's zones. *Chapter 1*.

optical emission spectroscopy (OES) Reading the light a plasma gives off at chosen wavelengths, to follow its chemistry or to detect an etch endpoint. *Chapter 7*.

over-etch Etching time beyond the nominal clearing time, so that the slowest point of the wafer clears too. *Chapter 1*.

overlay How well a printed layer lines up with the layer beneath it. *Chapter 4*.

override control A structure in which a second controller takes over or trims the action when a limit is approached. *Chapter 2*.

pad conditioning Roughening a polishing pad with an abrasive disk so that it keeps removing material at a steady rate. *Chapter 11*.

pad glazing The smoothing of a polishing pad's

surface with use, which lowers the removal rate unless the pad is conditioned. *Chapter 11*.

PID control Control by proportional, integral and derivative action on the difference between setpoint and measurement. *Chapter 1*.

plasma etching Removing a film with reactive gas and ions from an electrically driven plasma, which can etch straight down. *Chapter 1*.

polysilicon Very pure silicon made of many small crystals, the raw material from which single crystals are pulled. *Chapter 2*.

post-exposure bake The bake after exposure that drives the chemistry of a chemically amplified resist and so sets the linewidth. *Chapter 5*.

Preston equation The rule that polishing removal is proportional to the pressure times the relative speed of pad and surface. *Chapter 11*.

pull drive The motor that lifts the growing crystal; its speed is the fast actuator of the diameter loop. *Chapter 3*.

pyrometer An instrument that reads temperature from the radiation of a surface without touching it. *Chapter 1*.

rapid thermal processing (RTP) Heating one wafer at a time with lamps, quickly and briefly, instead of a batch in a furnace. *Chapter 9*.

recipe The settings a tool runs for one step on a wafer or lot, such as times, powers, pressures and doses. *Chapter 1*.

resist A light-sensitive film spread on the wafer, in which exposure, bake and development open the pattern. *Chapter 5*.

RF match network A circuit between a radio-frequency generator and the chamber that tunes itself so the power goes into the plasma. *Chapter 7*.

run-to-run control (R2R) Adjusting a tool's recipe between runs, lot by lot or wafer by wafer, from measurements of earlier runs. *Chapter 1*.

self-bias voltage The negative voltage that an electrode driven at radio frequency takes on, which sets how hard ions strike the wafer. *Chapter 7*.

sheet resistance The resistance of a thin layer per square, measured with a probe, which shows how well its dopants were activated. *Chapter 9*.

Siemens reactor A reactor that grows polysilicon from trichlorosilane on thin silicon rods heated by their own electric current. *Chapter 2*.

slip Crystal damage in which planes of atoms shift, caused by the stress of a temperature difference across a hot wafer. *Chapter 9*.

spatial modes Patterns of power or temperature across a plate that respond independently, each with its own gain and time constant. *Chapter 1*.

spike anneal A rapid thermal anneal that ramps the wafer quickly to a high temperature, holds it only briefly and lets it cool at once. *Chapter 9*.

structure limit The pull rate above which the crystal stops growing as one crystal; no instrument measures it and it falls as the crystal fattens. *Chapter 3*.

supervisory control A layer that watches the loops below and changes their setpoints, modes or permissions instead of moving actuators itself. *Chapter 8*.

thermal budget The heat a wafer receives in an anneal, counted as equivalent seconds at a reference temperature. *Chapter 9*.

thermal runaway A rise in temperature that feeds on itself, as when a hotter rod conducts better and draws more power. *Chapter 2*.

thermocouple A temperature sensor made of two wires of dissimilar metals, whose junction gives a voltage that rises with temperature. *Chapter 1*.

trichlorosilane A silicon compound, a gas at reactor conditions, from which pure silicon is deposited on the heated rods. *Chapter 2*.

virtual metrology Estimating a wafer's property from the tool's own sensor data, between real measurements. *Chapter 1*.

void An empty space left inside a plated trench when copper closes the top before it has filled the bottom. *Chapter 10*.

wafer A thin disk sliced from a silicon crystal, on which chips are built layer by layer. *Chapter 1*.

wafer bow The curvature of a wafer that keeps parts of it off the bake plate, so that those parts bake cooler. *Chapter 5*.

wet clean Maintenance in which an etch chamber's walls are cleaned with liquids, after which the chamber must season again. *Chapter 6*.

within-wafer nonuniformity The spread of a film's thickness, or of any process result, across one wafer. *Chapter 11*.

Index of Control Topics

AI agents
  completion checks, 298
  engineering practice, 296
  helper checks, 298
  open problems, 300
  reproduction, 298
  truth audit, 297
  verification record, 297
  while studying, 17
analyzer
  cyclic voltammetric stripping, 258
  laboratory practice, 259
  sample line and analysis delay, 249
  sampled and late, 241
Arrhenius kinetics, 34
Arrhenius rate, 114

back-calculation, 34, 118, 145, 189, 222
batch process
  no steady state, 52
batch sequencing, 170
Bayesian estimation, 275
bumpless transfer, 16, 118, 147, 169, 191

cascade control, 188
  crystal puller, 75
  outer trim, 198
chance constraint, 274
constraint
  unmeasured process constraint, 57

data export, 17
dead time
  in series with a lag, 95
decentralized control, 9, 109
decoupling, 119, 178, 212
derivative action
  noise gain, 69
  on a double integrator, 60
design of experiments
  D-optimal, 276
digital twin, 295
discrete integral control, 145
disturbance estimation, 66
disturbance model
  output bias, 171
double EWMA, 140
double integrator, 60

emissivity, 6, 215
endpoint detection, 10, 165, 293
  small open area, 182
estimator reset, 276
EWMA controller, 6, 144
  double, 274

fault detection
  coincidence rule, 198
  specificity, 202
fault diagnosis, 124
fault-tolerant control, 191
feedforward
  amp-hour, 240
  between process steps, 11, 145, 274, 291
    history, 146
  inverse model, 212
  known event, 119, 172
  model-based, 94
  preview, 82
film thickness measurement, 282
final control element, 12
flow scheme, 4

gain mismatch, 147
gain scheduling, 23, 220
  closed-form schedule, 33
  exact schedule, 61
  in practice, 45
  scheduling variable, 33
graded challenge, 41, 154, 204, 229, 281
graph Laplacian, 114
gross-error detection, 250
group delay, 91

hybrid systems, 198

inferential control, 6, 23
  as a cross-check, 41
  open problems, 46
  unmeasured constraint, 34
  wafer stage, 103
influence matrix, 269

information form, 275
innovation, 275
integral action
    shared measurement, 150
integral hold, 120, 197
integrator windup, 189
interaction, 221
ion source
    lifetime, 206
iterative learning control, 104

Kalman filter, 16, 66, 225
    steady-state, 171

least squares
    grid fit, 99

machine learning
    polishing, 295
maintenance event, 276
mass-transfer limitation, 29
measurement bias, 35, 94, 143, 221
measurement validation, 188, 189
    fast pre-gate, 199
metrology, 5
metrology delay, 7, 137, 272
metrology sampling, 8, 99, 143, 272, 295
mid-ranging control, 54
    history, 55
    loop separation, 62
model predictive control, 170, 224
    against simpler control, 294
    history, 215
    in the apps, 16
    move blocking, 172
    on an identified model, 170
    open problems, 76
    shrinking horizon, 35
    two actuators, 65
motion control
    commissioning, 102
moving horizon estimation, 16
multivariable control, 163, 273
multizone thermal control, 9, 109, 212, 293

outlier gate, 250, 275
over-etch, 10, 175
override control, 35

PID control
    anti-windup, 15, 34, 169
    sampled at analyses, 250
    velocity form, 15, 34, 118, 169, 221, 250
PID tuning
    SIMC rule, 123, 170
predictive maintenance, 295
profile control, 265
purge
    make-up balance, 248
pyrometry, 6, 215, 293
    calibration, 234
    lightpipe radiation thermometer, 233

ratio control
    dose to load, 249
reachable subspace, 278
recipe optimization, 163, 273
regularization, 273
relative gain array, 110, 118, 166
    triangular plant, 166
retry limit, 203
run-to-run control, 6, 112, 137, 139, 182, 224, 265
    design procedure, 293
    EWMA, 251
    history, 3, 140, 268, 295
    in practice, 130, 157
    in the fab, 12
    maintenance events, 300
    matched-lot estimator, 7, 149
    metrology delay, 7, 147, 291
    noise and drift, 7, 149
    open problems, 298
    pad age, 274
    per zone, 120
    profile, 273

sampled-data control, 240
    stability limit, 252
scenarios, 17
    completion rules, 17
scorecard, 17
selector, 26
self-regulation, 32
sensor bias, 127
sensor drift, 98
Siemens reactor
    history, 26
simulation engine, 15
simulator framework, 14
soft sensor, 26
spatial filter, 120
spatial frequency, 86

spatial frequency response, 116
spatial modes, 9, 116, 271
spectral radius, 252
statistical floor, 98
supervisory control, 198, 275
    bath life, 251

thermal runaway, 32
time-scale separation, 164
transport delay, 82
trim loop, 93

valve position control, 54
virtual metrology, 6, 272, 293, 294
    open problems, 284

Index of Applications

across-wafer uniformity, 168
additives
    accelerator, 241
    leveler, 241
    suppressor, 241
alignment marks, 88
amp-hour counter, 249
analyzer magnet, 191
arc chamber, 191
arc storm, 203
azimuthal nonuniformity, 272

bake plate, 109
    hundred-zone, 110
beam scan, 195
bell jar, 26
bleed and feed, 248
boron, 30
boron trifluoride, 191
Bossung curve, 85, 141
breakdown products, 243
    measurement, 260
bright-ring camera, 69
    in practice, 74

carrier head, 268
CD-SEM, 140
    in practice, 156
chamber seasoning, 142, 166
chemical mechanical planarization, 265
chemically amplified resist, 110
    history, 111
Child-Langmuir law, 191
clearing transition, 168
consumable wear, 266
copper electroplating, 240
critical dimension, 88, 137
    radial signature, 126
crucible lift, 55
cyclic voltammetric stripping, 241
    history, 242
Czochralski crystal growth, 52

damascene copper, 241
    history, 243
defect band, 70
deposition rate, 29
depth of focus, 95
dopant activation, 218
dose ledger, 195
dose monitor, 98
dose repaint, 203
dump-and-remake, 248

edge ring, 217
electric heater
    time-proportioned, 130
electro-pneumatic transducer, 283
etch recipe
    in practice, 181
excimer laser, 85
exposure dose, 85
exposure latitude, 141

Faraday cup
    in practice, 204
Faraday's law, 245
fill window, 246
first-wafer effect, 168
flash lamp annealing, 295
fluidized-bed reactor, 25
four-point probe, 251
Freeman source, 190
    history, 191

glitch
    acceptance test, 206

halogen lamp, 217
    in practice, 233
heater failure, 124
high-voltage glitch, 195

in-line metrology, 5
indirectly heated cathode, 190
    power supplies, 205
instrument, 12
integrated metrology, 140
ion implantation, 188
    history, 189
ion source, 188
ion-assisted etching, 165

lens heating, 85
  model feedforward, 100
level sensor, 83
  history, 85
  in practice, 100
leveling, 94
lithography and etch line, 137
lithography scanner, 82

mask selectivity, 168
meniscus, 55
metering pump, 259
metrology
  three-dimensional devices, 158
millisecond annealing, 235

optical emission spectroscopy, 170
  in practice, 180
overlay, 88

pad break-in, 284
pad conditioning, 270
pad glazing, 270
plasma etching, 10, 163
  history, 166
  inductively coupled chamber, 164
plating bath, 240
  additive analysis, 295
polisher, 265
polishing pad, 266
polysilicon, 23
post-exposure bake, 109, 142
power supply
  current or voltage regulation, 28
  rod heating, 45
Preston equation, 270
  history, 267
pull drive, 75
pyrometer, 26
  radiation, 44

qualification wafer, 275

radial temperature profile, 28
rapid thermal processing, 111, 212
  history, 214
recipe
  as final element, 156
resist
  stochastic effects, 132
RF match network, 174
  in practice, 180

self-bias voltage, 168
semiconductor manufacturing, 1, 163
  history, 3
  milestones, 2
sheet resistance, 218
sidewall angle, 168
Siemens reactor, 23
sight glass, 40
silicon
  metallurgical grade, 25
  resistivity, 26
silicon crystal puller, 52
site flatness, 85
slip (silicon wafer), 214
slurry, 268
source deposits, 195
space-charge limit, 191
spike anneal, 212
stabilization step, 173
step-and-scan exposure, 82
stepper, 84
structure limit, 57

thermal budget, 218
thermal dose, 114
thermal stress, 214
thermionic emission, 192
thermocouple, 115
  in practice, 129
trichlorosilane, 25

virgin make-up solution, 243
void, 246

wafer bow, 114
wafer flatness, 85
  warpage, 103
wafer flow, 4
wafer stage, 86
  in practice, 101
wafers per hour, 89
wall polymer coverage, 168
wet clean, 142
within-wafer nonuniformity, 271

About the Editor

John D. Hedengren is a professor in the Department of Chemical and Biological Engineering at Brigham Young University, where he leads the Process Research and Intelligent Systems Modeling (PRISM) group. He received the B.S. and M.S. degrees in chemical engineering from Brigham Young University and the Ph.D. degree from the University of Texas at Austin in 2005. He developed advanced process control applications for polymer production at ExxonMobil Chemical Company and commissioned nonlinear model predictive controllers with PAS, Inc. before joining the faculty at Brigham Young University in 2011.

His teaching covers process dynamics and control, engineering computing, data science, machine learning, and dynamic optimization, with course material published online for use inside and outside the classroom. He developed the Temperature Control Lab, a portable heat transfer device now used by many universities for process control and data science instruction, and he maintains the APMonitor and GEKKO optimization software used in those courses. He serves as Chair of Control Education for the IEEE Control Systems Society and as a trustee of the CACHE Corporation. He received the AIChE David Himmelblau Award in 2014, the AIChE Computing Practice Award in 2018, and the John R. Ragazzini Award in 2024, and he was inducted into the Process Automation Hall of Fame in 2025.

Volumes in This Series

Interactive Control Engineering: Principles and Applications has 20 volumes. Nine are organized by control technology and eleven by industry or application area. Every chapter in every volume has a simulator at https://apmonitor.com/apps/.

1 **Process Dynamics, Modeling, and Identification.** How plants respond, and how to find the model.
2 **Practical PID Control.** Tuning, anti-windup and implementation that works on real loops.
3 **Feedforward, Cascade, and Ratio Control.** Choosing a structure from where the disturbance enters.
4 **Override, Split-Range, and Constraint Control.** When limits decide what the controller may do.
5 **Delays, Slow Measurements, and State Estimation.** Controlling what arrives late or cannot be measured.
6 **Multivariable and Plantwide Control.** Pairing loops, managing inventories, closing the recycle.
7 **Nonlinear Dynamics, Oscillations, and Stabilization.** Plants that will not stay put without control.
8 **Batch, Run-to-Run, and Hybrid Control.** Recipes, endpoints and learning from the last run.
9 **Model Predictive Control and Economic Optimization.** Planning moves ahead, and choosing where to operate.
10 **Control in Upstream and Midstream Oil and Gas.** From the wellbore to the terminal.
11 **Control in Petroleum Refining.** Throughput, quality and limits from crude unit to blending.
12 **Control in Petrochemicals and Polymers.** Severity, selectivity and grade from cracker to pellet.
13 **Control in Chemical Plants.** Recycle, equilibrium and energy in continuous chemical processes.
14 **Control in Semiconductor Manufacturing.** Run-to-run control where little can be measured in the chamber.
15 **Control in Mining and Mineral Processing.** Variable ore, late assays and equipment at its limit.
16 **Control in Iron, Steel, and Metals Production.** Slow furnaces, fast strip, and measurements that arrive once.
17 **Control of Energy Conversion and Storage.** Fast regulation and slow dispatch on a changing grid.
18 **Control of Nuclear and Fusion Systems.** Reactor physics, protection limits and coordinated control.
19 **Control of Biological Systems and Human Performance.** Variable subjects, sparse samples and limits that cannot be crossed.
20 **Motion Control, Robotics, and Autonomous Systems.** Tracking, navigation, contact and allocation.